An Introduction to Calculus and Algebra

VOLUME 2

Calculus Applied

An Introduction to Calculus and Algebra

Volume 2 *Calculus Applied*

An Introduction to Calculus and Algebra

Volume 2

Calculus Applied

The Open University Press

Contents

Contents

Contents

NOTE

References to particular examples or exercises are made throughout by
giving chapter, section, and example or exercise number; Example 4 in
Chapter 1 Section 1 would thus be referred to as Example 1.1.4, and so on.

Editors' Preface

This is the second of three volumes presenting some of the essential concepts of mathematics, a few important proofs (usually in outline), together with exercises designed to reinforce the understanding of the concepts and to develop the beginning of technical skill.

The major part of the material used here has been selected from the correspondence texts of the *Open University Foundation Course in Mathematics*. Open University courses provide a method of study, at university level, for independent learners, through an integrated teaching system which includes textual material, radio and television programmes local tutorial arrangements, and short residential courses. The correspondence text components of the Mathematics Foundation Course were produced by a Course Team (the names of the members are listed below) and were edited by Professor M. Bruckheimer and Dr. Joan Aldous.

The selection of material for these three volumes has been made with the needs of students of other subjects particularly in mind, by the Course Team preparing the second-year short course *Elementary Mathematics for Science and Technology*, and the three volumes constitute the set book around which the course is designed. (The members of this Course Team are listed below.)

In preparing these volumes, the Course Team has attempted to provide the kind of mathematics which is particularly useful for students who already have some knowledge of Science or Technology, but who, before proceeding in their own subjects, need to deepen their appreciation of the mathematical concepts underlying the techniques of calculus and algebra.

The special character of the original Foundation Course texts has been preserved as far as is possible, but the scope of the volumes is narrower than that of the course, which endeavours to give an overall picture of mathematics. For a much fuller appreciation of what mathematics is and what mathematics does, the reader is therefore referred to the original Mathematics Foundation Course correspondence texts.

WALTON 1971

GRAHAM FLEGG
ROGER MEETHAM

Mathematics Foundation Course Team

Professor M. Bruckheimer
Dr. J. M. Aldous
Mr. D. J. A. Burrows
Mr. R. Clamp
Mr. S. N. Collings
Dr. A. Crilly
Dr. D. A. Dubin
Mr. H. G. Flegg
Mr. E. Goldwyn
Mr. N. W. Gowar
Dr. A. Graham
Mr. R. D. Harrison
Mr. H. Hoggan

Mr. F. C. Holroyd
Miss V. King
Mr. R. J. Knight
Dr. J. H. Mason
Mr. R. Nelson
Miss J. Nunn
Professor R. M. Pengelly
Professor O. Penrose
Dr. G. A. Read
Mr. J. Richmond
Mr. E. Smith
Professor R. C. Smith

Course Assistants

Mr. J. E. Baker
Mr. W. D. Crowe

General Course Consultant

Mr. D. E. Mansfield

Elementary Mathematics for Science and Technology Course Team

Professor R. M. Pengelly
Mr. H. G. Flegg
Dr. A. R. Meetham
Mr. L. Aleeson
Mr. G. Burt
Dr. J. K. Cannell
Mr. R. Clamp
Dr. P. M. Clark
Mr. S. N. Collings

Dr. A. Cooper
Dr. A. Crilly
Professor M. J. L. Hussey
Mr. E. G. Law
Mr. F. B. Lovis
Mrs. V. Richards
Dr. R. A. Ross
Dr. T. B. Smith

Course Assistant

Mr. R. W. Duke

Notation

x

CHAPTER 1 STATIONARY VALUES OF FUNCTIONS OF ONE VARIABLE

1.0 Introduction

Many problems in both pure and applied mathematics are concerned with maximum or minimum properties of some sort. For example, at what angle should a missile be fired in order to give the maximum range? What is the largest area which can be surrounded by a given length of fencing? What is the shortest path between two points on a given surface? Problems of this kind are sometimes called *optimization problems*, and some of them can be attacked systematically using calculus.

In this chapter we discuss functions of one real variable, by which we mean *real functions* (whose domains and codomains are R or subsets of R). In Chapter 8 of Volume 1, we introduced the concept of the *derivative* of a real function. We now start by seeing how we can use this concept to develop techniques for determining maxima and minima.

In order to understand the argument which we shall be using and to attempt the exercises, you need to have a sound grasp of the principles introduced in Volume 1 and to know the standard derived functions which were given there.

1.1 Using the Derivative

To introduce a method of optimization which uses the derivative, let us first look at a fairly simple function.

Example 1

What are the greatest and least values of the images of the function f, where $f : x \longmapsto x^3 - 7x^2 + 8x + 1 \qquad (x \in [0, 5])$?

Graph of
$x \longmapsto x^3 - 7x^2 + 8x + 1$
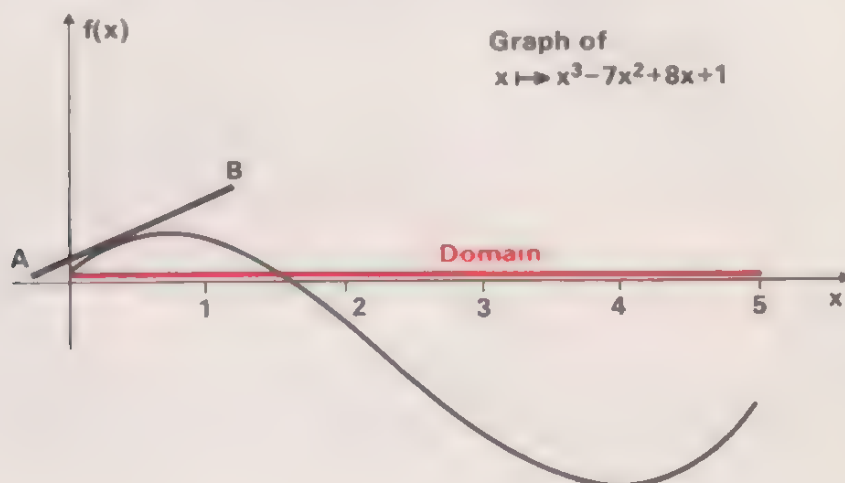
Imagine the tangent line AB moving along the curve from the point where $x = 0$ to the point where $x = 5$. The slope of this line is initially positive, becomes negative, and is positive again when we reach $x = 5$. At two intermediate points the line is parallel to the x-axis (it has zero slope) and the graph shows that these are the points at which $f(x)$ takes its greatest and least values in the interval $[0, 5]$. Remember that $f'(x)$ is the slope of the tangent at x. If

$$f(x) = x^3 - 7x^2 + 8x + 1 \qquad (x \in [0, 5]),$$

then we know that the slope at x is given by

$$f'(x) = 3x^2 - 14x + 8.$$

The values of x for which $f'(x) = 0$ are the two solutions of the quadratic equation:

$$3x^2 - 14x + 8 = 0,$$

which are

$$x = \tfrac{2}{3} \quad \text{and} \quad x = 4.$$

The greatest and least values of $f(x)$ in the interval $[0, 5]$ are therefore

$$f(\tfrac{2}{3}) = 3\tfrac{14}{27} \quad \text{and} \quad f(4) = -15$$

respectively.
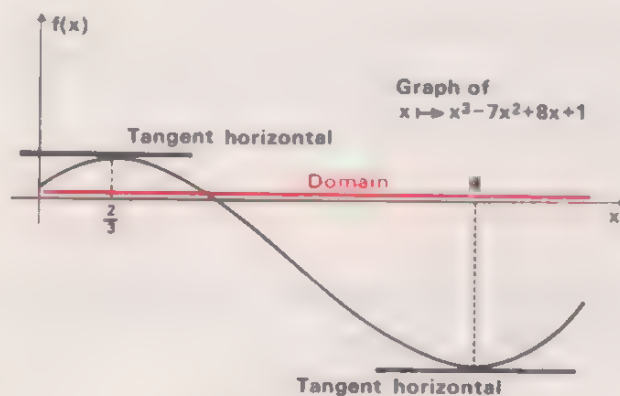
Exercise 1

(i) Find the greatest value of $g(x)$, where

$$g: x \longmapsto 4 - x^2 \qquad (x \in [-2, 2]).$$

(ii) Find the greatest rectangular area which can be enclosed by a fence of length 100 metres.

This technique has to be used with a certain amount of care on some occasions. The following examples illustrate the sort of difficulties which can occur.

Example 2

What is the greatest value of

$$f(x) = x^3 - 7x^2 + 8x + 1 \qquad \text{in } [0, 6]?$$

You may well say that the answer is $3\frac{14}{27}$, as in Example 1. But, if so, how do you explain the fact that $f(6) = 13$?

The apparent contradiction is explained when we examine the graph of the function, which shows that the greatest value of $f(x)$ in $[0, 6]$ occurs when $x = 6$.

Graph of
$x \longmapsto x^3 - 7x^2 + 8x + 1$

Domain
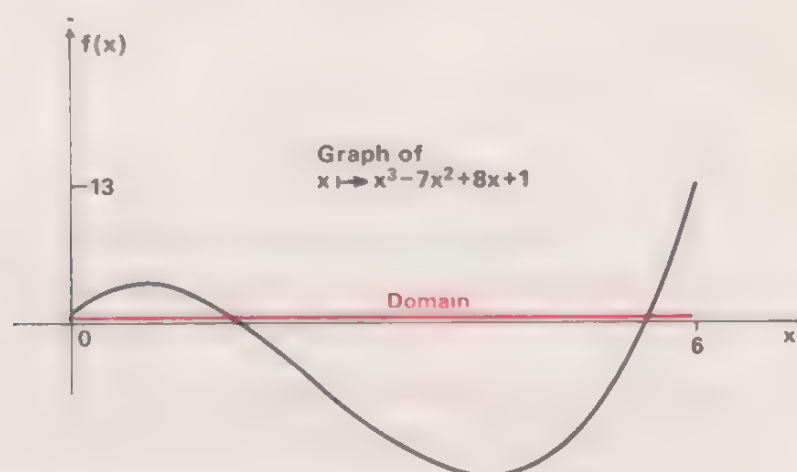
Example 3

What is the least value of $f(x)$, where

$$f : x \longmapsto |x| \qquad (x \in R)?$$

This function (which you met in Volume I) is called the *modulus function* it has the following graph:

Graph of
$x \longmapsto |x|$

3
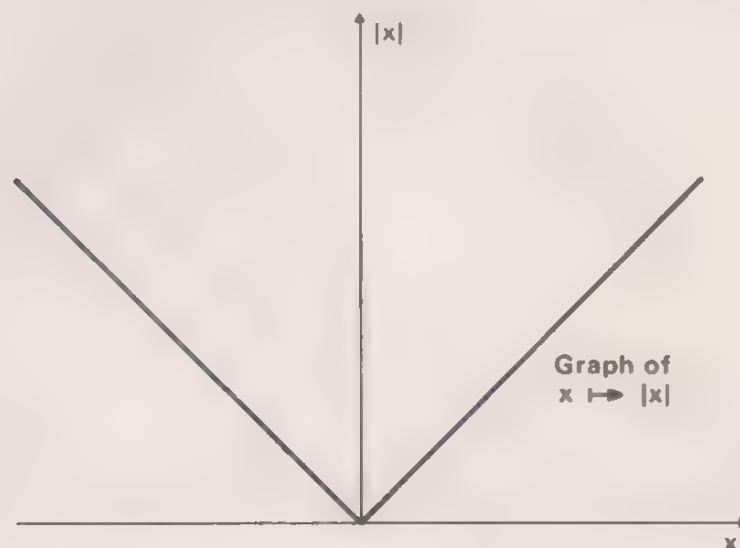
The difficulty in this case is that f is not differentiable for all values of x. We saw in Chapter 8, Volume I that

$$f'(x) = +1 \qquad \text{if } x > 0,$$

and

$$f'(x) = -1 \qquad \text{if } x < 0;$$

but if $x = 0$, then the limit:

$$\lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

does not exist, and so f is not differentiable at $x = 0$.

Unfortunately in this case it is this very point which is of most interest, for we can see from the graph that it is at this point that $f(x)$ takes its least value.

Sometimes it is impossible to specify the greatest value of the images of a function, simply because there is no such value. Consider the function:

$$f : x \longmapsto x^2 \qquad (x \in R).$$

In this case, there are always elements in the domain for which $f(x)$ takes values greater than any fixed number you care to name. (Notice that we may not say "the greatest value is infinity" because infinity is *not* a number, and by the words "greatest value" we imply that we are looking for a number.)

The following strategy will enable us to deal with most problems without having to draw graphs. (It is nearly always useful, however, to draw a rough sketch of the graph when this is not too difficult.)

This strategy will enable us to find the greatest and least values of the images of a well-behaved* function whose domain is the interval $[a, b]$.

---

* By a "well-behaved" function we mean a continuous function which is differentiable at all but a *finite* number of points in its domain. It *is* possible to find a function which is continuous at all points in its domain, but not differentiable at *any* point in its domain, and to which the above strategy does not apply. (For the definition of a continuous function, see Volume 1, Chapter 6.)

```
┌─────────────────────────────┐
│   Given the problem of finding │
│   the greatest and least values │
│   of f(x) where f is continuous and │
│        has domain [a,b]         │
└─────────────────────────────┘
```

```
┌─────────────────────────┐          ┌─────────────────────────┐
│  Find the values of x other │          │   Find the values of x    │
│ than a and b for which f'(x) │          │    for which f'(x) = 0    │
│        does not exist       │          │                           │
└─────────────────────────┘          └─────────────────────────┘
```

```
┌─────────────────────────────┐
│  Evaluate f(x) at these values of x │
└─────────────────────────────┘
```

```
┌─────────────────────┐
│  Evaluate f(a) and f(b) │
└─────────────────────┘
```

```
┌─────────────────────────┐          ┌─────────────────────────┐
│   Take the largest of all   │          │    Take the least of all    │
│    these values of f(x)     │          │     these values of f(x)    │
│        This is the          │          │         This is the         │
│     GREATEST VALUE          │          │       LEAST VALUE           │
└─────────────────────────┘          └─────────────────────────┘
```
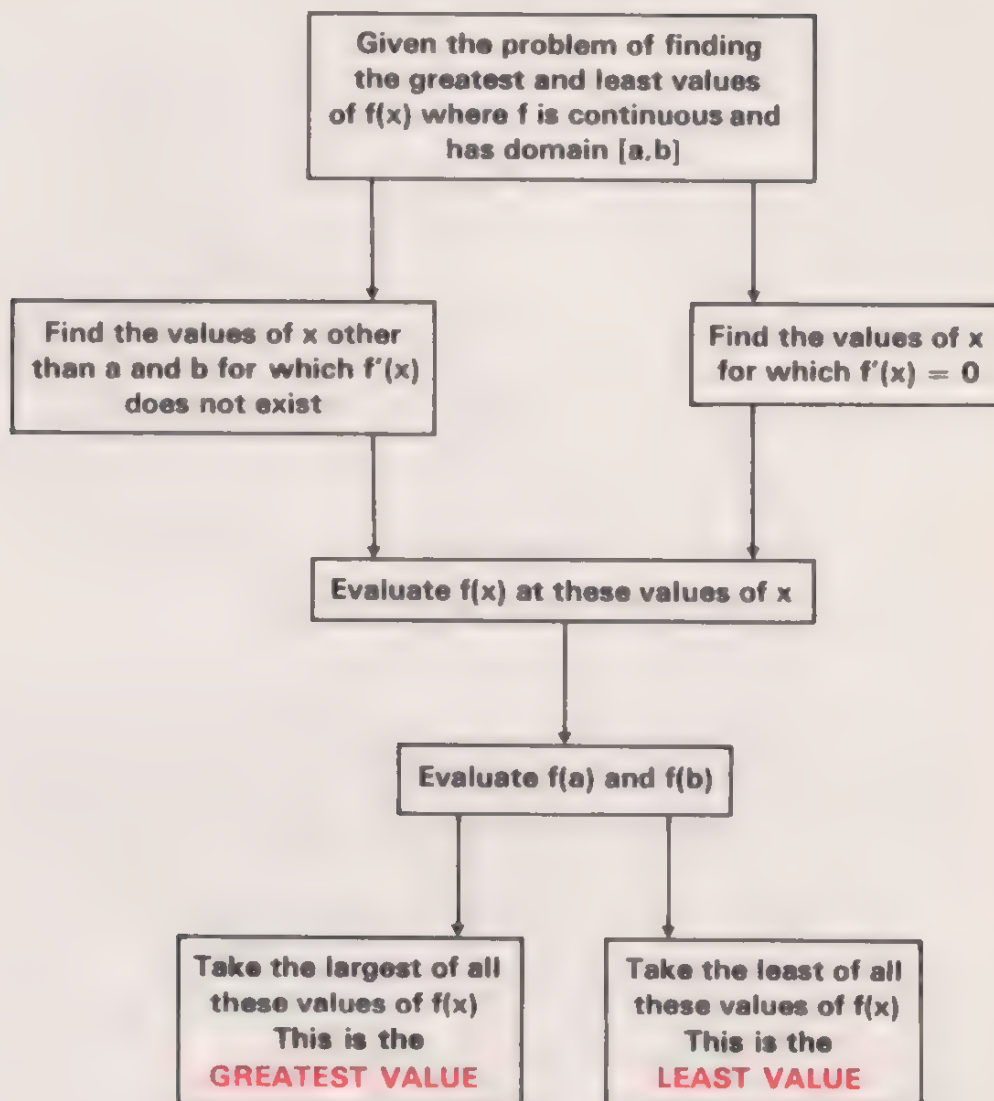
Exercise 2

Find the greatest or least value (as appropriate) of the images of the function

$$f : x \longmapsto x + \frac{1}{x} \qquad (x \in R^{+})$$

and sketch the graph of f.

Some More Set Notation

We introduce some set notation which will be useful in the following sections.

The Intersection of Two sets

To be a member of the set $\{x : x - 1 < 0$ and $x : x + 2 > 0\}$ an element must satisfy *both* conditions, or in other words it must belong to both of the sets

$$\{x : x - 1 < 0\} \quad \text{and} \quad \{x : x + 2 > 0\}$$

$$\{x : x-1 < 0\}$$

overlap

$$-2 \qquad -1 \qquad 0 \qquad 1$$

$$\{x : x+2 > 0\}$$
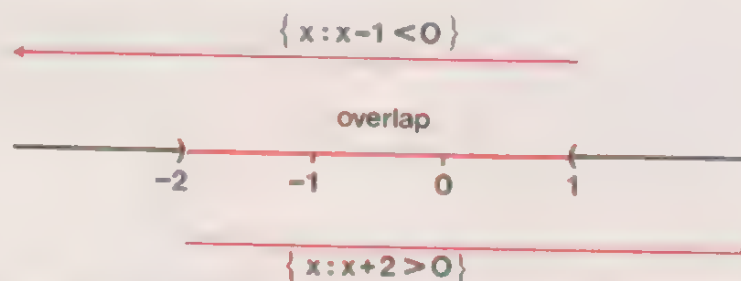
This "overlap" of two sets we call the intersection set.

In general we define the intersection of two sets A and B as the set of all elements which belong both to A *and* to B. We denote this set by $A \cap B$, which we read as "A intersection B".

This definition applies to any two sets A and B, but our number line illustration applies only to sets of real numbers.

As another illustration, consider two sets of points in a plane, each of which is enclosed by a curve, and denoted in the diagram by A and B respectively.

The intersection of these two sets is the shaded region:

$$A \qquad\qquad B$$

$$A \cap B$$
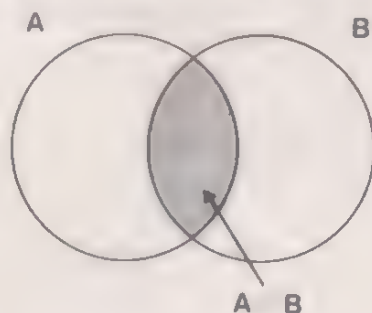
It is often convenient to use this type of diagram to illustrate more general sets. For example the region A could represent the set of all men and B the set of all University students. The shaded area would then represent the set $A \cap B$ which, in this case, would be the set of all male University students.

Exercise 3

Illustrate the following sets

  (i) $A = \{(x, y) : x \in R, y \in R, x + y - 1 = 0\}$.
 (ii) $B = \{(x, y) : x \in R, y \in R, x^2 + y^2 - 1 = 0\}$.
(iii) $A \cap B$.

1.2   Local Maxima and Minima

Although we often wish to find the greatest (or least) value of the images of a function, there are occasions when it is useful to be able to find points (like $x = \frac{2}{3}$ in Example 1.1.2) where there is a sort of "minor peak" on the "side of the mountain"; this is after all a first step towards the greatest value. In order to be more precise, we formulate the following definitions.

Let f be a given real function with domain A; let $c \in A$.

If there is a positive number ε such that

$$f(x) \leqslant f(c) \quad \text{for} \quad \text{all} \quad x \in A \cap [c - \varepsilon, c + \varepsilon]$$

then we say that f has a local maximum at c^*.

If there is a positive number ε such that

$$f(x) \geqslant f(c) \quad \text{for all} \quad x \in A \cap [c - \varepsilon, c + \varepsilon]$$

then we say that f has a local minimum at c.

To distinguish these values from the "greatest" and "least" values that we have been discussing, we shall call the greatest (or least) value taken by the images of a function its overall maximum (or minimum).

How do the above definitions apply in the context of Example 1.1.2? The function:

$$f : x \longmapsto x^3 - 7x^2 + 8x + 1 \qquad (x \in [0, 6])$$

has a local maximum at $\frac{2}{3}$. If we take $\varepsilon = \frac{1}{3}$, say, then $f(x) \leqslant f(\frac{2}{3})$ for all $x \in [\frac{1}{3}, 1]$.

We shall also say, for example, "$(c, f(c))$ (or simply "c") is a local maximum", when it would be more correct to say "$(c, f(c))$ is a local maximum point".

Graph of
f : x ⟼ x³−7x²+8x+1
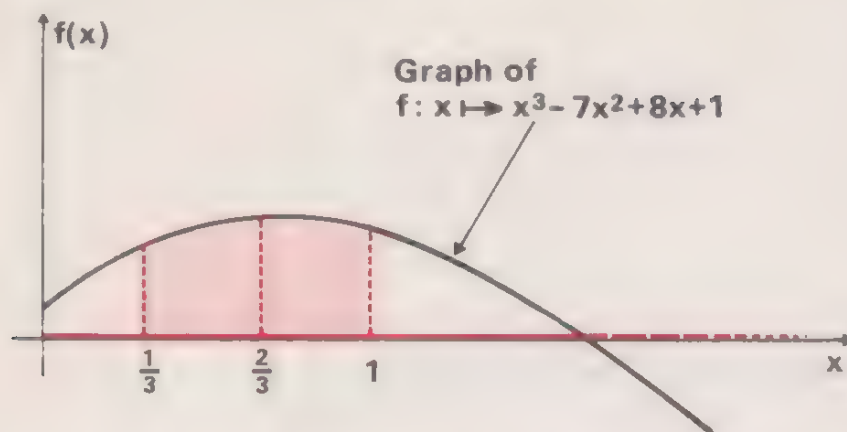
This function also has a local maximum at 6, because for all x in $A = [0, 6]$ close to 6 we have $f(x) \leqslant f(6)$. For example, taking $\varepsilon = 0.1$, the set $A \cap [c - \varepsilon, c + \varepsilon]$ becomes $[5.9, 6]$, and $f(x) \leqslant f(6)$ in this interval.

Graph of
f : x ⟼ x³−7x²+8x+1
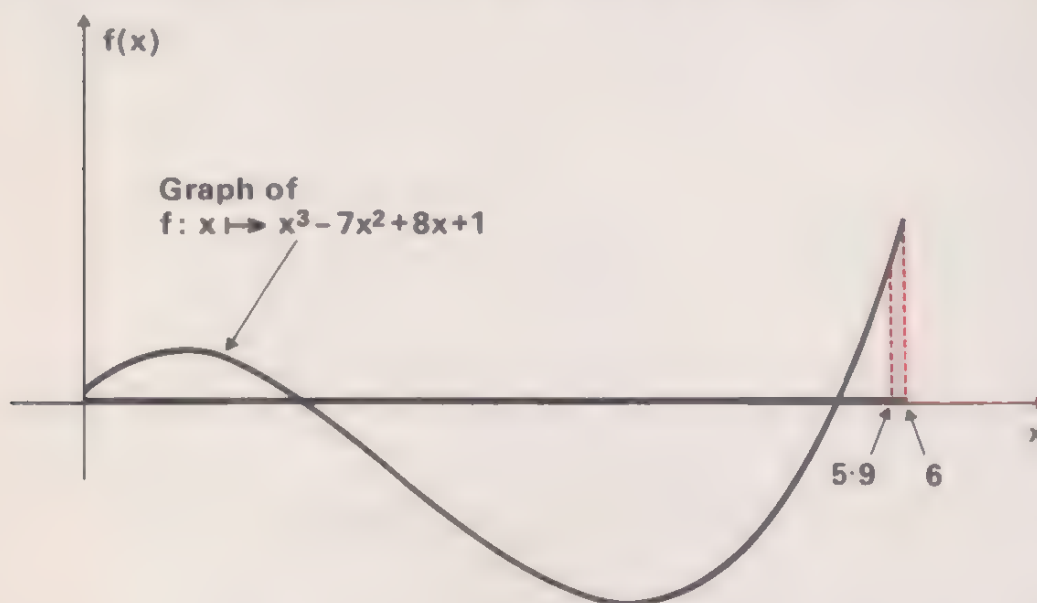
In this case we know that $f(6)$ is in fact the overall maximum value of $f(x)$ for $x \in A$. The reason for taking $A \cap [c - \varepsilon, c + \varepsilon]$ in our definition is that we are only interested in that part of the interval $[c - \varepsilon, c + \varepsilon]$ which lies in A. If c is not an end-point of the domain, we are interested in the behaviour of f close to c *on both sides* of c: but, for example, if c is at the right-hand end of the domain, we need only look to the left of c.

Effectively, we are saying that a point which is a local maximum is an overall maximum in its immediate surroundings, and similarly for a local minimum. Speaking very roughly, if it rains on the graph of a function, the puddles collect around the local minima, and the water runs away to the overall minimum when the puddles overflow. A local maximum would be a suitable place to stand in a flood, but an overall maximum would be preferable.

Of course, an overall minimum is also a local minimum: but a local minimum need not be an overall minimum. The following graphs show some of the various possibilities. Each of them corresponds to some function f.
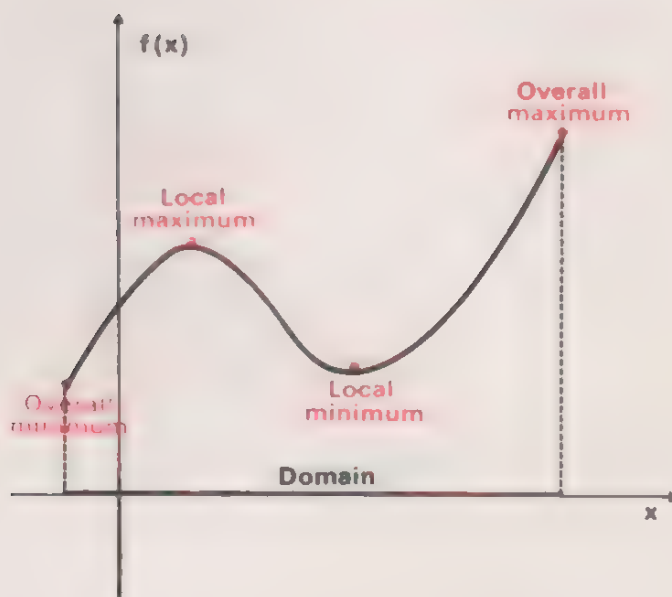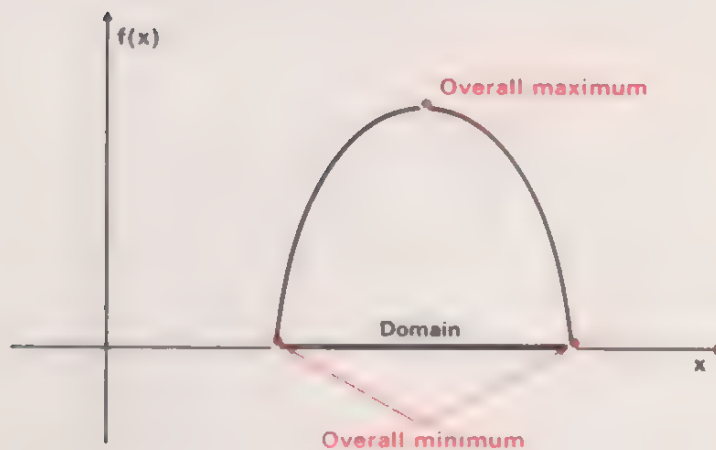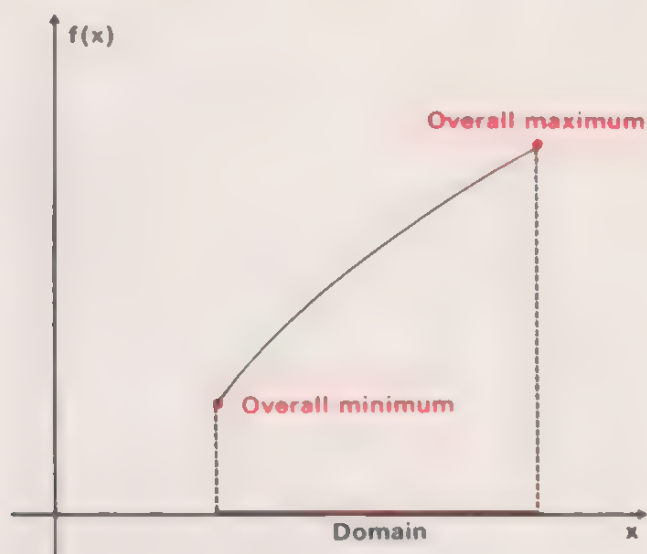
Exercise 1

Mark the local and overall maxima and minima on the following graphs:

(i)
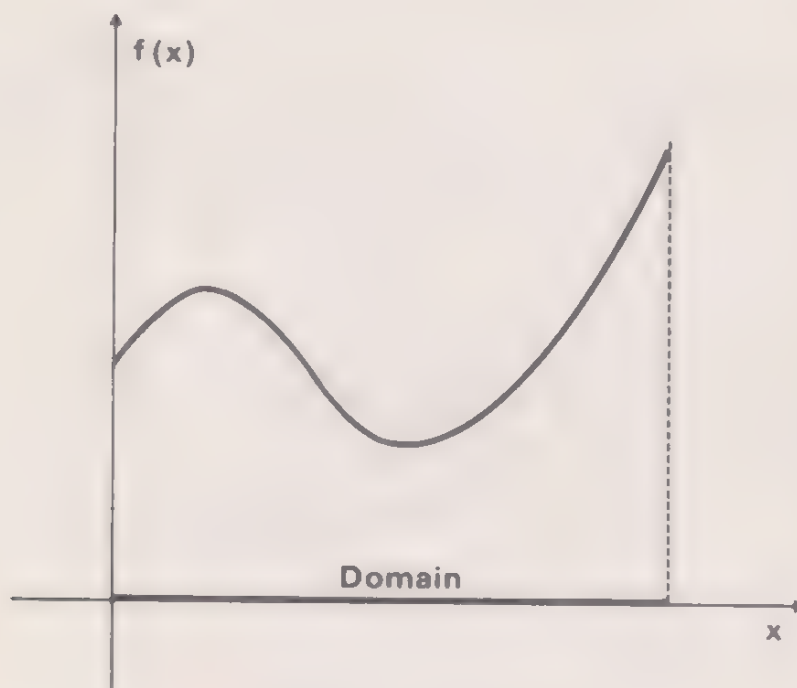
(ii)
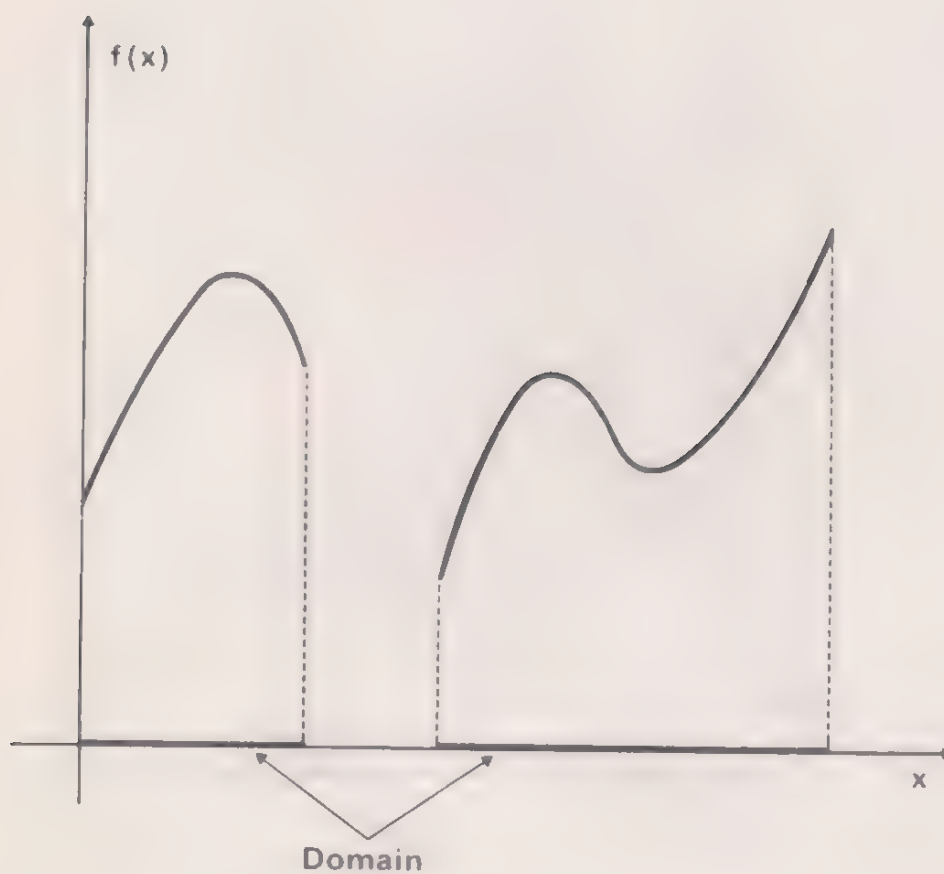
We call points x such that $f'(x) = 0$ stationary points of f. A stationary point is thus simply a point on the x-axis where the tangent at the corresponding point on the graph is parallel to the x-axis.

If we wish to locate local maxima (or minima) of a function, it would seem a sound idea to first locate the stationary points. However, there are unfortunately two complications.

We have already seen that a local maximum (or minimum) of a function can occur at a point which is not a stationary point (in other words where the slope of the graph is not zero) either because the function is not differentiable at that point, and "slope" is meaningless, or because the point occurs at an end-point of the domain. We shall overcome this difficulty by considering only functions which are differentiable at all points of their domains, and by restricting our search for the local maxima and minima of such a function to points which are not end-points of the domain, and then examining the end-points as a separate issue.

There is a second complication which is more serious: a stationary point may be neither a local maximum nor a local minimum, as we shall see in the next example.

Example 1

Consider the function

$$f : x \longmapsto x^3 \qquad (x \in R)$$

which has the graph:

Graph of $x \longmapsto x^3$
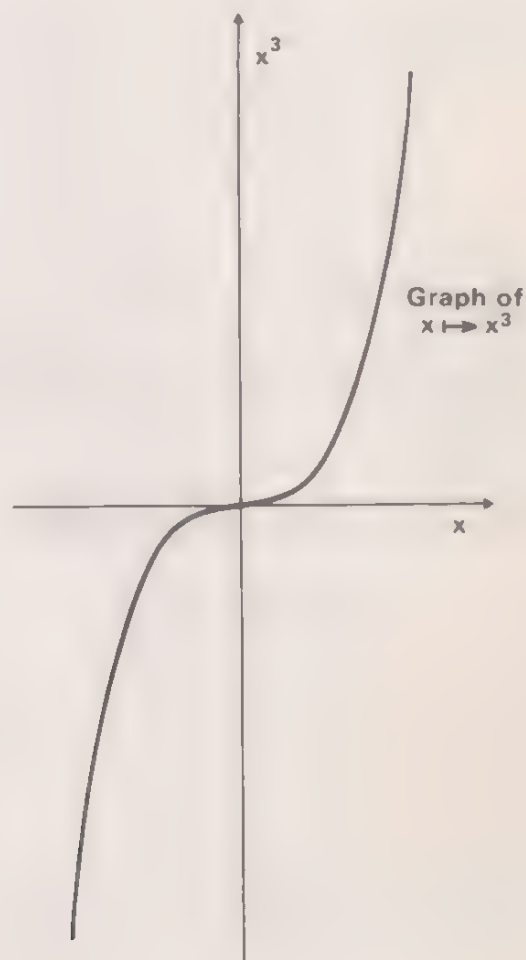
We know that

$$f' : x \longmapsto 3x^2 \qquad (x \in R),$$

so that

$$f'(0) = 0,$$

and therefore f has a stationary point at 0. But f has neither a local maximum nor a local minimum at 0.

How then can we distinguish between local maxima, local minima and stationary points which are neither?

1.3 Two Useful Methods

In this section we shall describe two methods for determining the nature of stationary points.

Method One

Let f be a real differentiable function:

$$f : x \longmapsto y \qquad (x \in A).$$

If $f'(x) < 0$ for all $x \in S$, where $S \subseteq A$, then we say that $f(x)$ is decreasing in S.

If $f'(x) > 0$ for all $x \in S$, where $S \subseteq A$, then we say that $f(x)$ is increasing in S.

If there is a point $c \in R$ such that

$$f'(x) < 0 \quad \text{if} \quad x < c,$$

and

$$f'(x) > 0 \quad \text{if} \quad x > c,$$

then $f(x)$ is decreasing in $\{x : x < c\}$ and increasing in $\{x : x > c\}$. Clearly, $f(c)$ is the *overall minimum* of $f(x)$.
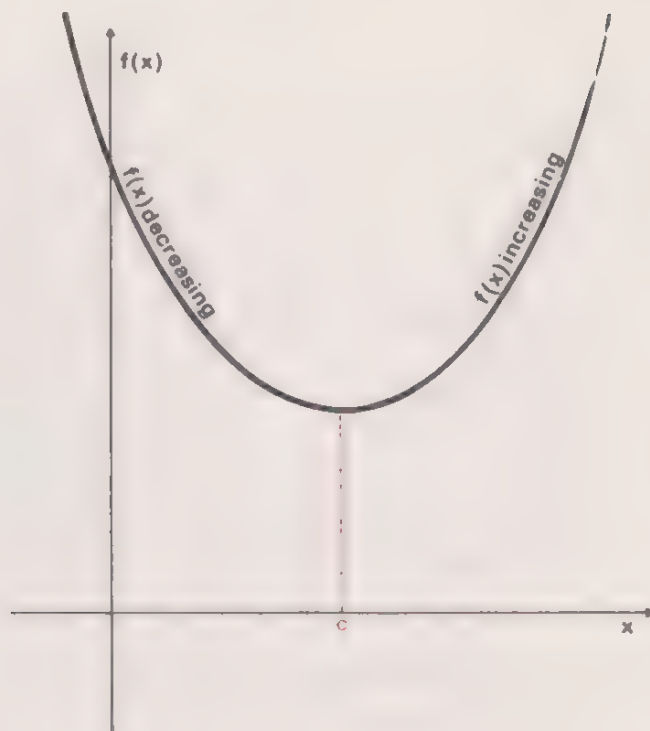
If there is a point $c \in R$ and a positive number ε such that

$$f'(x) < 0 \quad \text{if} \quad c - \varepsilon < x < c,$$

and

$$f'(x) > 0 \quad \text{if} \quad c < x < c + \varepsilon,$$

then, although we cannot say anything about overall maxima and minima, we can be sure that $f(c)$ is a *local minimum* of $f(x)$.
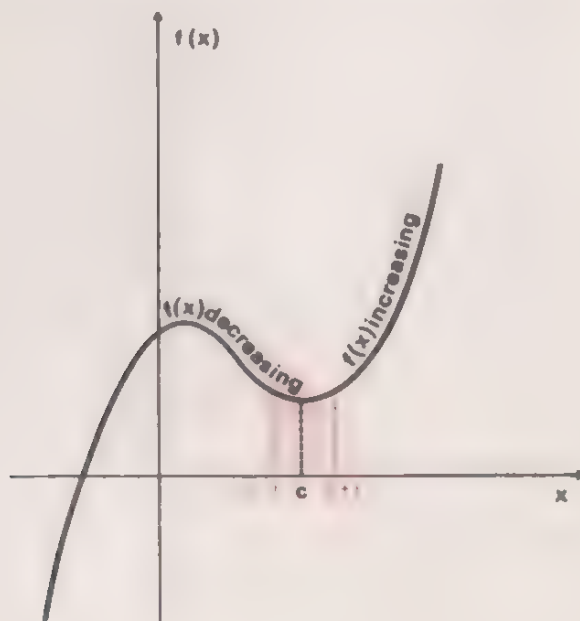
Method Two

You may be familiar with another technique which involves the second derived function of the function under consideration. This technique does, however, require a certain amount of care, not so much in its application, which is often very straightforward, but in the conclusions which you draw. Unfortunately, most students entirely ignore the method which we have just discussed once they are introduced to this second technique.

We know that the derived function, f', can be used to study the rate at which the images, $f(x)$, of a function f are changing. Similarly, f'' can be used to study the rate of change of $f'(x)$. In our previous method it was the fact that $f'(x)$ changed sign at a local maximum or minimum which was important, and it is a short step to the question: "Can we predict a change in the sign of $f'(x)$, if we are given f''?"

To find an answer to this question, suppose that we are given a real continuous differentiable function:

$$g : x \longmapsto g(x) \qquad (x \in R),$$

and that g' is also a real continuous function. (We shall say in a moment how g is related to f.) Suppose further that $g(c) = 0$ and $g'(c) > 0$ for some $c \in R$; then, since g' is continuous, there must be an interval $[c - \varepsilon, c + \varepsilon]$ in which $g'(x) > 0$. It follows that $g(x)$ is increasing for $x \in [c - \varepsilon, c + \varepsilon]$. (As shown in the figure below, $g(x)$ may be increasing outside this interval as well, but that is immaterial to the argument.)
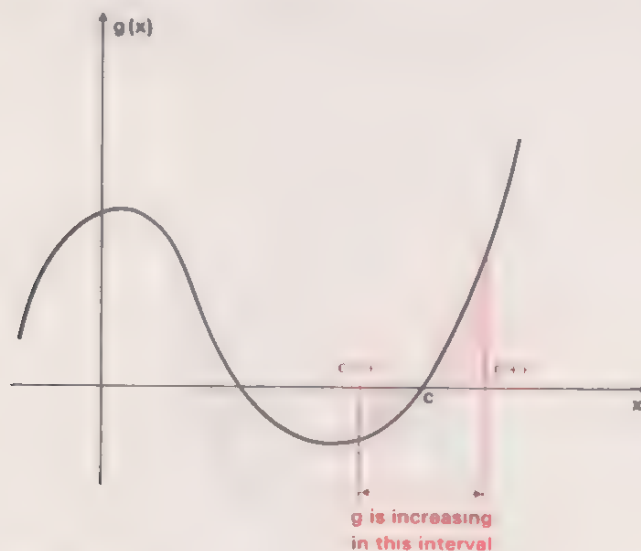
We see from the graph that

$$g(x) < g(c) \qquad \text{if} \quad c - \varepsilon < x < c,$$

and

$$g(x) > g(c) \qquad \text{if} \quad c < x < c + \varepsilon.$$

(Don't forget that one of our assumptions is that $g(c) = 0$.)

The useful piece of information which we are seeking follows from the above if we now assume that $g = f'$.

Notice that we require $g(c) = 0$, and this implies that

$$f'(c) = 0,$$

in other words, c is a stationary point of f. We also require that $g'(c) > 0$; since $g' = f''$, this implies that

$$f''(c) > 0.$$

With these conditions we were able to conclude that $g(x) < g(c)$ if $c - \varepsilon < x < c$, and $g(x) > g(c)$ if $c < x < c + \varepsilon$. Translated to give a result for f, this becomes

$$f'(x) < 0, \qquad \text{if } c - \varepsilon < x < c,$$

and $$f'(x) > 0, \qquad \text{if } c < x < c + \varepsilon.$$

This means that there is a *local minimum* at c.

If originally we had taken $f''(c) < 0$, then our final conclusion would be that there is a *local maximum* at c.

(We assumed for convenience that the domain of f was R, but the same results are true for any function which has an interval as its domain.)

Classification of Stationary Points Using the Second Derivative

What should be entered in the blank circle?
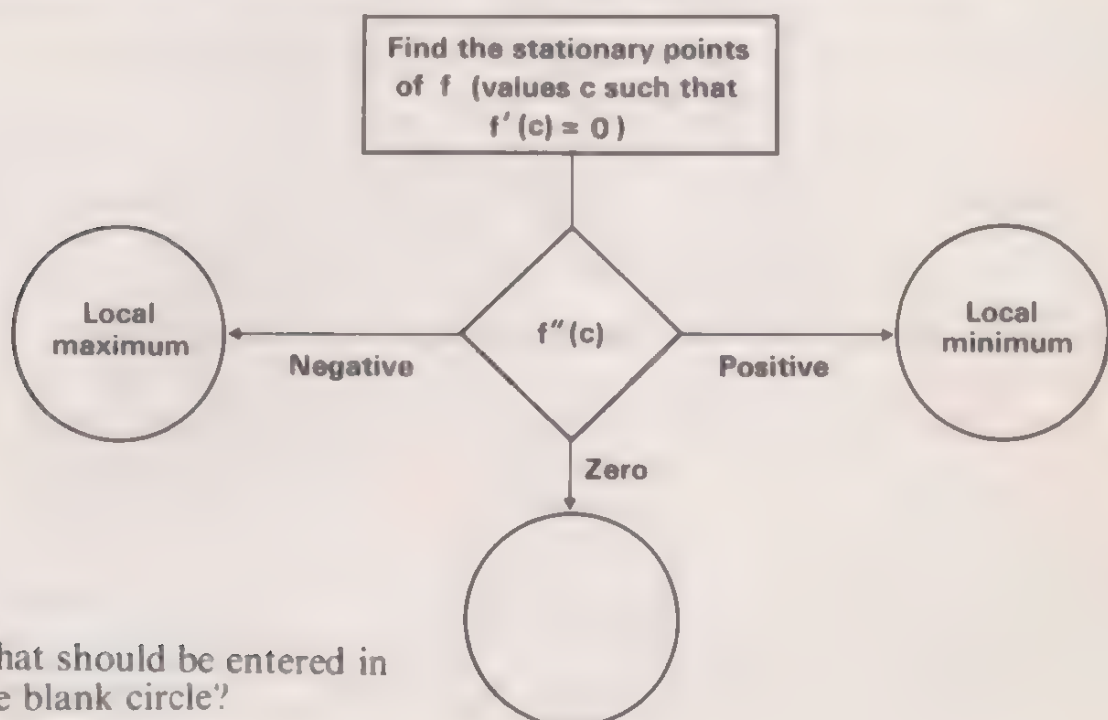
A Few Words of Warning

What can we say if $f'(c) = 0$ and $f''(c) = 0$? It is very tempting to say that f has neither a local maximum nor a local minimum at c, but this is wrong. The following examples should make the point clear.

Example 1

Consider the following three functions, each with domain R:

$$f : x \longmapsto x^4 \qquad\qquad g : x \longmapsto 1 - x^4 \qquad\qquad h : x \longmapsto x^3$$
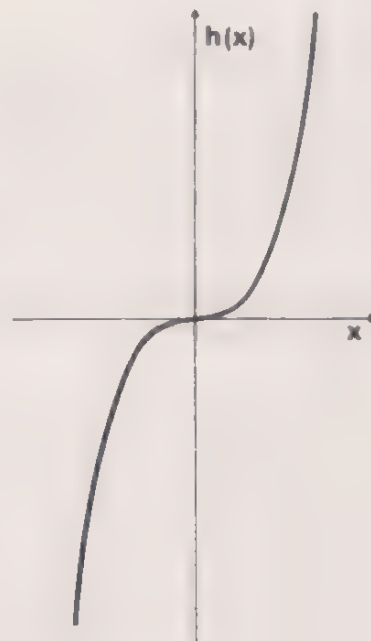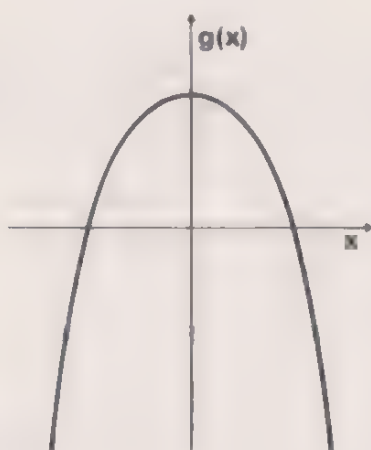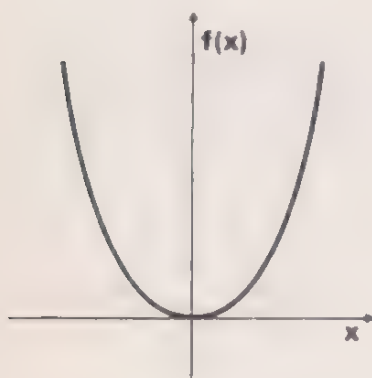
$f(x) = x^4$	$g(x) = 1 - x^4$	$h(x) = x^3$
$f'(x) = 4x^3$	$g'(x) = -4x^3$	$h'(x) = 3x^2$
$f'(0) = 0$	$g'(0) = 0$	$h'(0) = 0$
$f''(x) = 12x^2$	$g''(x) = -12x^2$	$h''(x) = 6x$
$f''(0) = 0$	$g''(0) = 0$	$h''(0) = 0$
From the graph, f has a local minimum at 0.	From the graph, g has a local maximum at 0.	From the graph, h has neither a local minimum nor a local maximum at 0.

What can we enter in the blank space on our flow chart corresponding to the case when $f'(c) = 0$ and $f''(c) = 0$? The answer is that the space is best left blank, for, given only this information, we can say nothing

except that the tangent to the graph at c is horizontal. To specify the nature of the stationary point, we require more information. There are more powerful tests for classifying stationary points using higher derivatives, but we leave these until later. (You may find it interesting to try to construct such a test for yourself.) Remember that Method One works even when $f''(c) = 0$.

Although calculus is a wonderful tool, it isn't a substitute for common sense. A little concentrated thought will occasionally go a long way, as you can see in the following example.

Example 2

Find the overall minimum value of

$$g(x) = ((x^4 + 2) + x^2(3 - x^2))^2 \qquad (x \in R).$$

If your first thought is: "Differentiate, and to the devil with the subtleties", then we admire your single-mindedness, but not your common sense.

The following solution is much simpler. Simplifying, we get

$$g(x) = (x^4 + 2 + 3x^2 - x^4)^2$$
$$= (2 + 3x^2)^2.$$

Since $x^2 \geqslant 0$, $2 + 3x^2$ takes its least value, 2, when $x = 0$; hence the overall minimum value of $g(x)$ is 4.

Exercise 1

Find the stationary points of the following functions, and classify each of them as a local maximum, a local minimum or neither.

(i) $f : x \longmapsto x^3 - 6x^2 + 9x + 6 \qquad (x \in R)$,
(ii) $h : x \longmapsto x \ln x \qquad\qquad\qquad (x \in R^+)$.

1.4 Additional Exercises

Exercise 1

$$f : x \longmapsto 3x^4 - 4x^3 \qquad (x \in [-2, 2])$$

(i) Determine the stationary points of f.
(ii) In which intervals is $f(x)$ increasing?
(iii) In which intervals is $f(x)$ decreasing?
(iv) What is the overall minimum value of f?
(v) What is the overall maximum value of f?

Exercise 2

Find the greatest or least value of the image of the function

$$g : x \longmapsto x^2 + \frac{16}{x} \qquad (x \in R^+).$$

Sketch the graph of g.

Exercise 3

Find the stationary points of the following functions and classify each of them as a local maximum, a local minimum or neither.

(i) $g : x \longmapsto 3x^4 - 4x^3 \qquad (x \in R)$

(ii) $S : x \longmapsto \sin x \qquad (x \in [-\pi, \pi]).$

1.5   Answers to Exercises

Section 1.1

Exercise 1

(i) Since $x^2 \geqslant 0$ for all x, the greatest image is $g(0) = 4$. To illustrate the method we are introducing, we also sketch the graph of g :
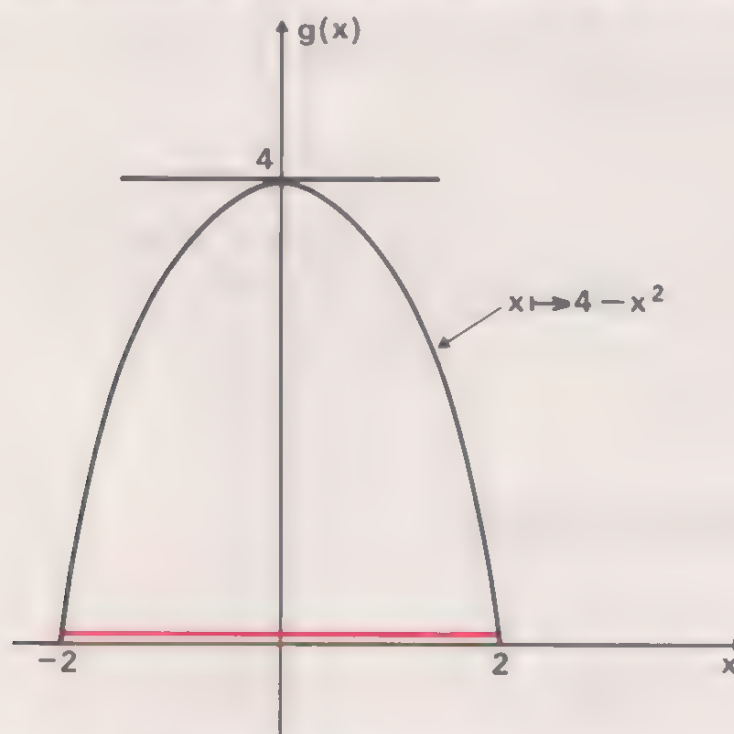
The slope of the tangent to the curve at the point x is

$$g'(x) = -2x$$

This slope is zero when $x = 0$. From the graph of g, we see that $g(0) = 4$ is the greatest of the set of images of f.

18

(ii)

x

50 − x
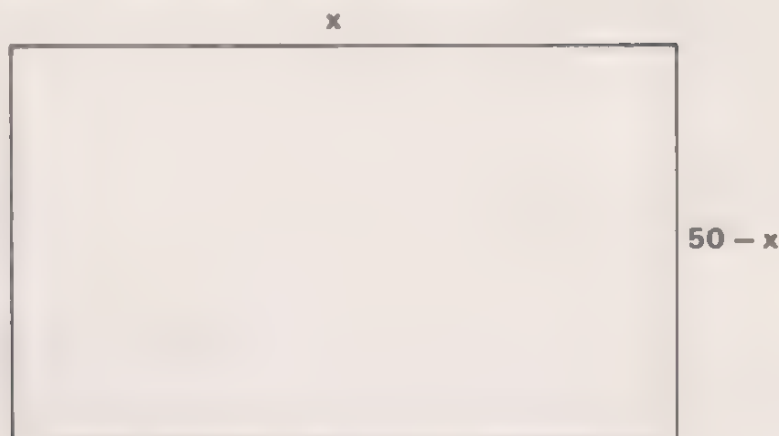
Let x be the length in metres of one side of the rectangle. The area of the rectangle is

$$x(50 - x)\, \text{m}^2,$$

so we can express the area of the rectangle by the function:

$$g : x \longmapsto x(50 - x) \qquad (x \in [0, 50]).$$

Then

$$g'(x) = 50 - 2x$$

and

$$g'(x) = 0 \text{ when } x = 25.$$

With this value for x, the rectangle is a square, and the required area is $625\,\text{m}^2$.

Exercise 2

$$f(x) = x + \frac{1}{x},$$

so that

$$f'(x) = 1 - \frac{1}{x^2}$$

and therefore

$$f'(x) = 0 \quad \text{if} \quad x = \pm 1.$$

However, only the value $x = 1$ is in the domain of f. On this occasion it helps to sketch a graph of f, and we can easily do this if we

notice the following:

(a) $x + \dfrac{1}{x}$ is never zero for $x \in R^+$;

(b) $x + \dfrac{1}{x}$ is very nearly equal to x when x is large, because the value

of $\dfrac{1}{x}$ is then very small;

(c) $x + \dfrac{1}{x}$ is very nearly equal to $\dfrac{1}{x}$ when x is small, because then

$\dfrac{1}{x}$ is large and predominates over x.

Using the above facts we can draw the following sketch:
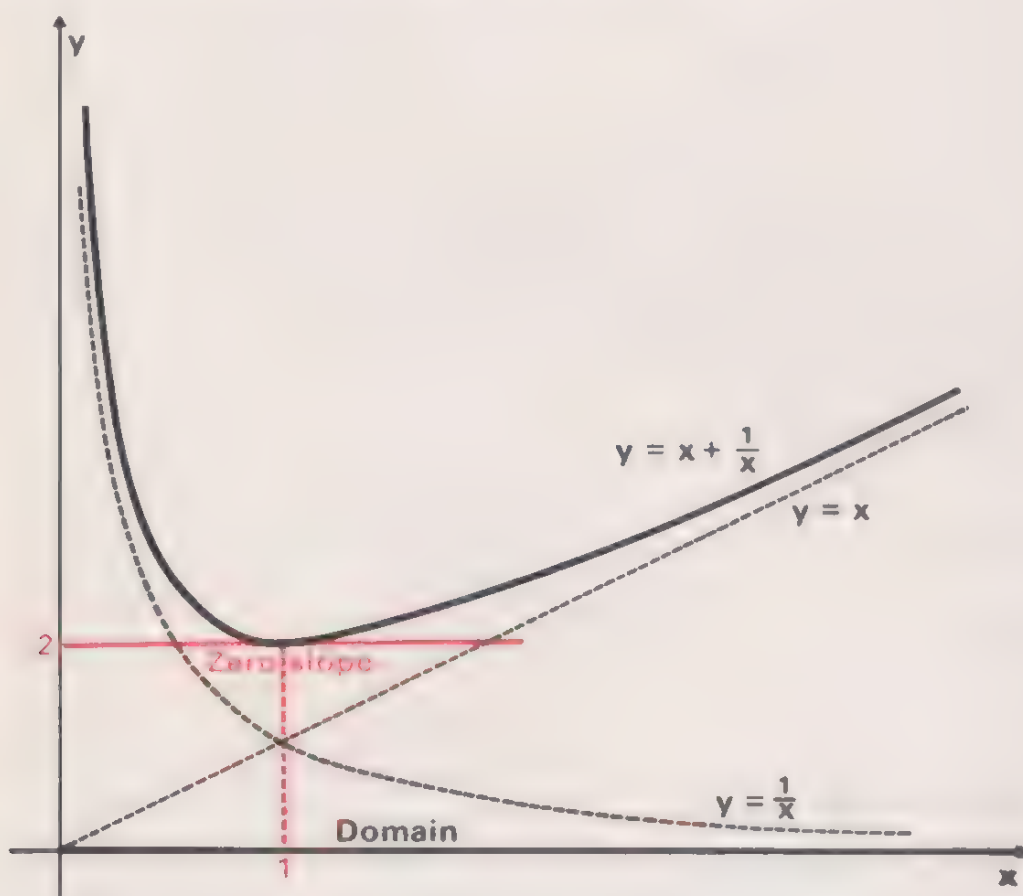
The least value of $f(x)$ is $f(1) = 2$; there is no greatest value.

Exercise 3

A ∩ B = {(0,1),(1,0)}
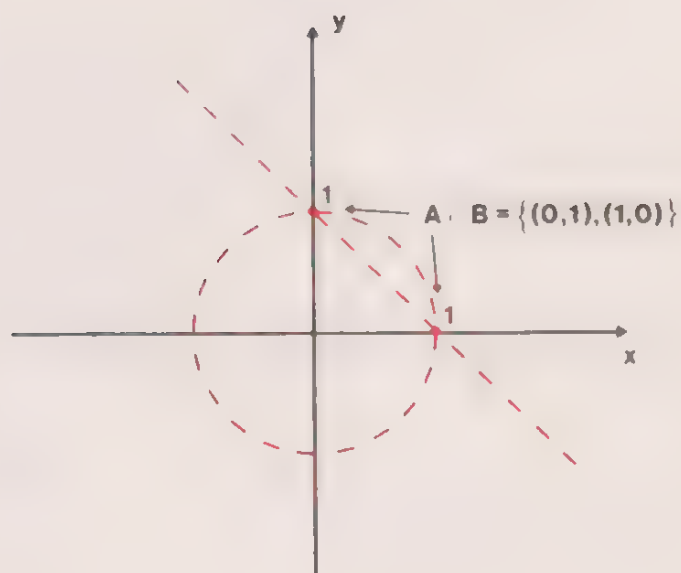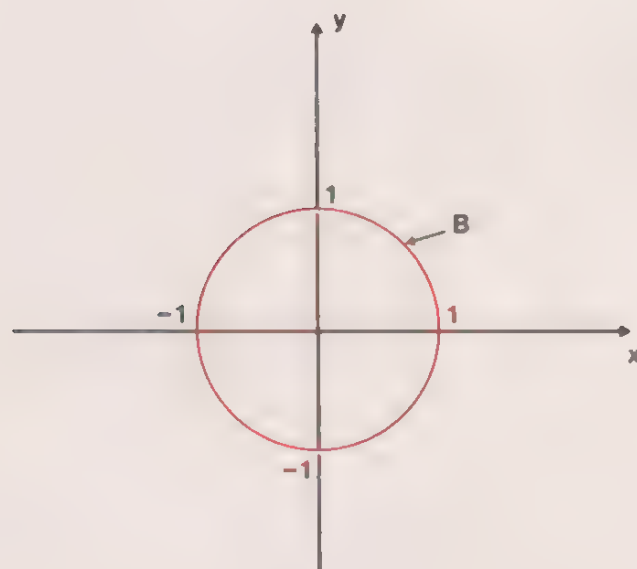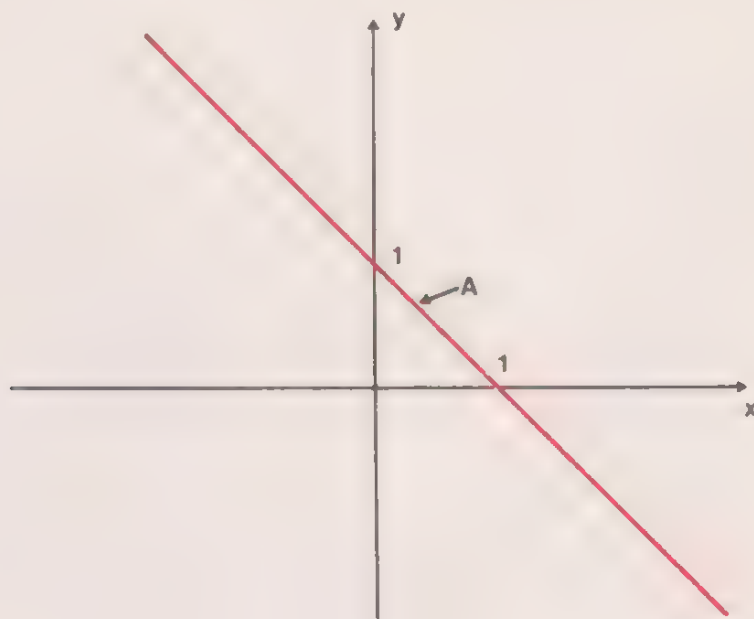
Section 1.2

Exercise 1

(i)
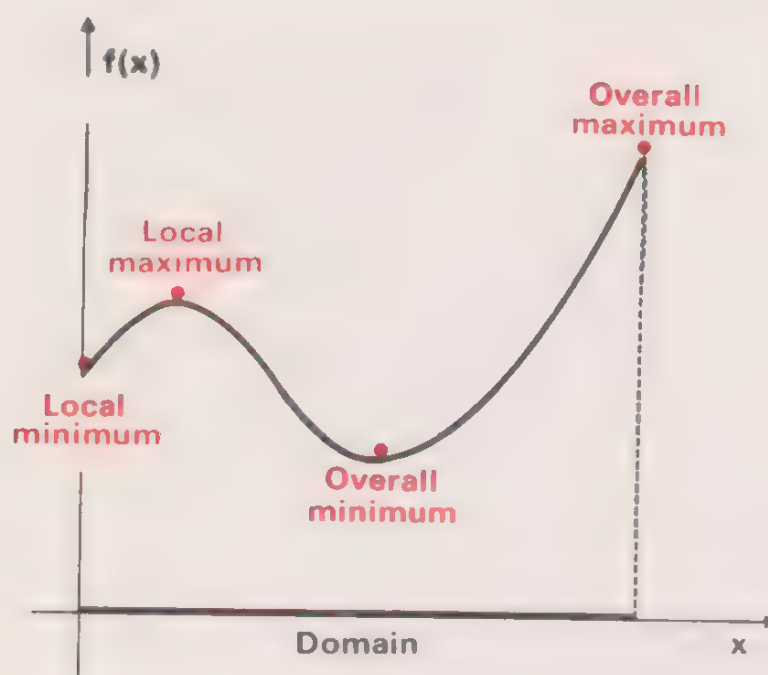
(ii)
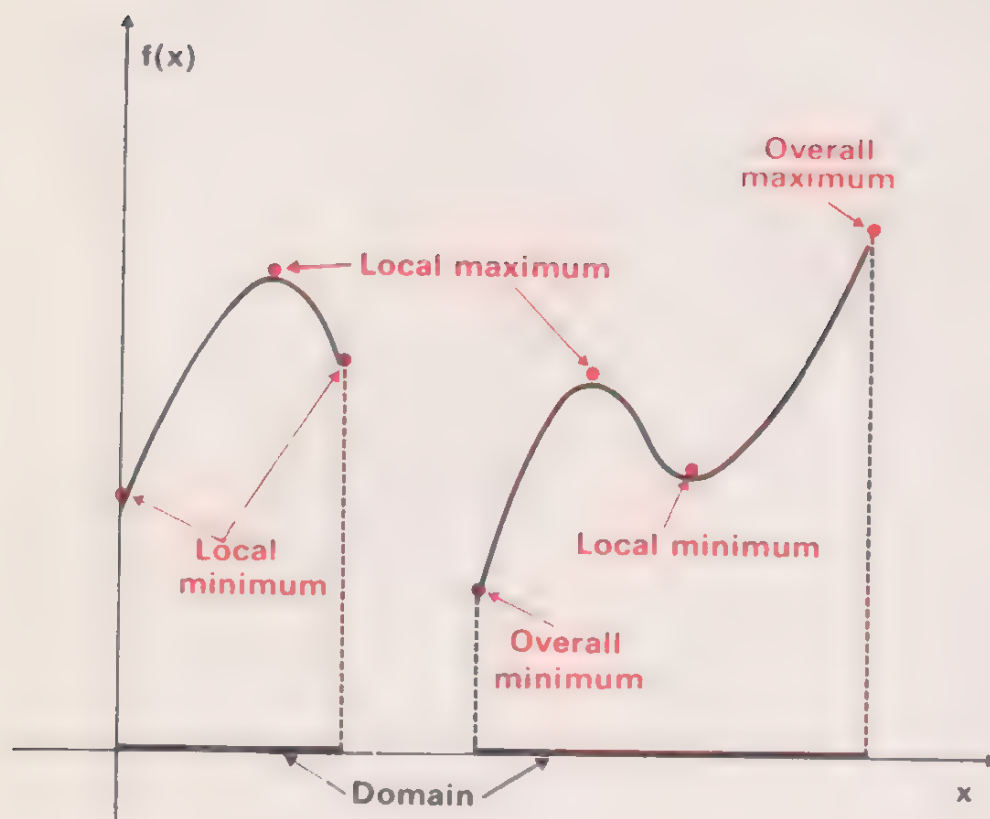
Section 1.3

Exercise 1

(i) $\qquad f'(x) = 3x^2 - 12x + 9$

$\qquad\qquad\qquad = 3(x - 3)(x - 1).$

Thus $f'(x) = 0$ when $x = 1$ and when $x = 3$.

$\qquad\qquad f''(x) = 6x - 12$

and

$\qquad\qquad f''(1) = -6,$

which is less than 0, giving us a local maximum at $x = 1$.

$\qquad\qquad f''(3) = 6,$

which is greater than 0, giving us a local minimum at $x = 3$.

(ii) $\qquad h'(x) = x \times \dfrac{1}{x} + \ln x$

$\qquad\qquad\qquad = 1 + \ln x$

Thus $h'(x) = 0$ when $\ln x = -1$, that is, when $x = \dfrac{1}{e}$.

$\qquad\qquad h''(x) = \dfrac{1}{x},$

which is greater than 0 when $x = \dfrac{1}{e}$, giving us a local minimum at this point; it is, in fact, an overall minimum.

Section 1.4

Exercise 1

(i) $\quad x = 0$ and $x = 1$

(ii) $\quad 1 < x \leqslant 2$

(iii) $\quad -2 < x < 0$ and $0 < x < 1$

(iv) $\quad f(1) = -1$

(v) $\quad f(-2) = 80$

Exercise 2

$$g(x) = x^2 + \frac{16}{x},$$

so that

$$g'(x) = 2x - \frac{16}{x^2},$$

and therefore

$g'(x) = 0$ if $2x^3 - 16 = 0$, that is, if $x = 2$.

Once again we can draw a sketch if we notice that:

(a) $x^2 + \dfrac{16}{x}$ is a little greater than x^2 when x is very large;

(b) $x^2 + \dfrac{16}{x}$ is a little greater than $\dfrac{1}{x}$ when x is very small (but not zero);

(c) $g(x) > 0$ when $x \in R^+$.

Using the above facts, we can sketch the following graph:
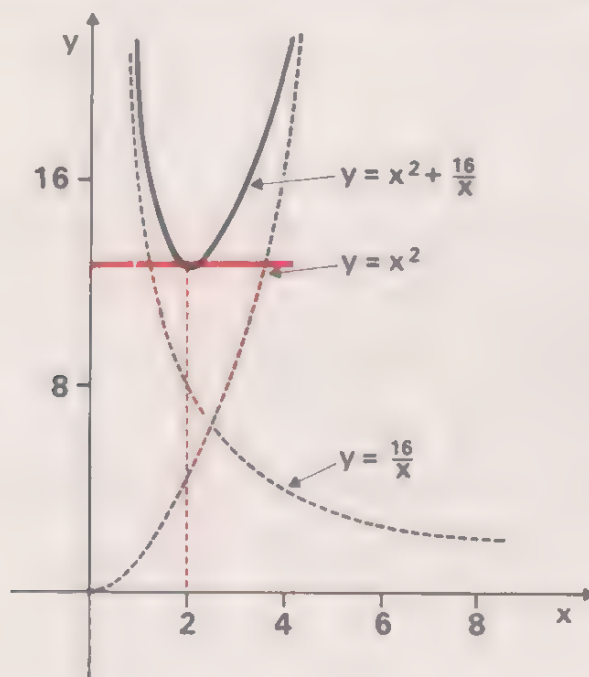
The least value of $g(x)$ is $g(2) = 12$; there is no greatest value.

Exercise 3

(i) $\qquad g'(x) = 12(x^3 - x^2) = 12x^2(x - 1),$

and

$\qquad g'(x) = 0$ if $x = 0$ or $x = 1,$

so these are the stationary points.

$$g''(x) = 12(3x^2 - 2x) = 12x(3x - 2),$$

so

$$g''(0) = 0 \text{ and } g''(1) = 12.$$

Immediately we can see that there is a local minimum at 1, but Method Two breaks down at 0 and we can obtain no information from it. However, our Method One will still work. We can see that $g'(x)$ can never be positive if $x < 1$, and so it is certainly negative for x near to and on either side of 0. It follows that g can have neither a local maximum nor a local minimum at 0.

(ii) $$S'(x) = \cos x,$$

and

$$S'(x) = 0 \text{ if } x = \pm\frac{\pi}{2},$$

so these are the stationary points,

$$S''(x) = -\sin x,$$

so

$$S''\left(-\frac{\pi}{2}\right) = 1, S''\left(\frac{\pi}{2}\right) = -1.$$

Hence we can deduce that S has a local maximum at $\frac{\pi}{2}$ and a local minimum at $-\frac{\pi}{2}$.

CHAPTER 2   FUNCTIONS OF TWO VARIABLES

2.0   Introduction

In this chapter we extend the principles which we have discussed in Chapter 1 to functions of two real variables. We consider the Cartesian space of three dimensions. A function of two variables often defines a surface in this space. In particular, we consider the general equation of a plane.

In order to be able to discuss stationary values of functions of two variables we need to extend the concept of *derivative* to such functions, and we do this intuitively before giving the formal definition of *partial derivatives*. Finally, we investigate stationary points in this new context.

2.1   Representation of Functions

We shall continue to investigate the problem of optimization, but now we shall concentrate on functions of two (real) variables; that is, functions of the form:

$$F : (x, y) \longmapsto z \qquad ((x, y) \in R \times R),$$

where $z \in R$. Before doing this we need to know a little three-dimensional co-ordinate geometry, because it is often helpful to represent such functions by surfaces. We really don't have enough time to do justice to geometry here; in fact a dedicated geometer would probably say that we were hardly doing geometry anyway. Our purpose in this section is to enable you to visualize the functions, and to describe the corresponding techniques in a pictorial and intuitive fashion. Later we shall apply these geometric notions to our problem of optimization.

We shall restrict the discussion to functions of two variables, although the basic results have equivalent forms for any number of variables. (A function of n real variables is a function which maps an element of the form $\underbrace{(x, y, \ldots, w)}_{n \text{ terms}}$ to a real number; that is, its domain is a subset of the Cartesian product, $\underbrace{R \times R \times \ldots \times R}_{n \text{ terms}}$ (which is usually denoted by R^n), and its codomain is R.)

We can represent many functions of one (real) variable by (pictorial) graphs, which enable us to use our intuition when examining the behaviour of the functions. In particular, when thinking of maxima and minima

of such functions, we find the graphical approach very helpful. You should notice, however, that we try to discard the purely pictorial arguments in favour of symbolic reasoning, as soon as we feel that we are on the right track.

We shall base a number of arguments on pictures because we think that they are easier to understand this way.

Our first thought is to find a diagram which represents a function of two (real) variables, rather as a graph represents a function of one (real) variable. In this sort of diagram, we shall find that a function can often be represented by a surface. All our functions will be assumed to be "well-behaved"; in other words, the surfaces representing them have no spikes, gaps, or similar oddities.

Cartesian Co-ordinates

In the Cartesian plane each ordered pair of real numbers (a, b) corresponds to a unique point.

Likewise in the Cartesian space of three dimensions, each ordered triple (x, y, z) of real numbers corresponds to a unique point of the space.
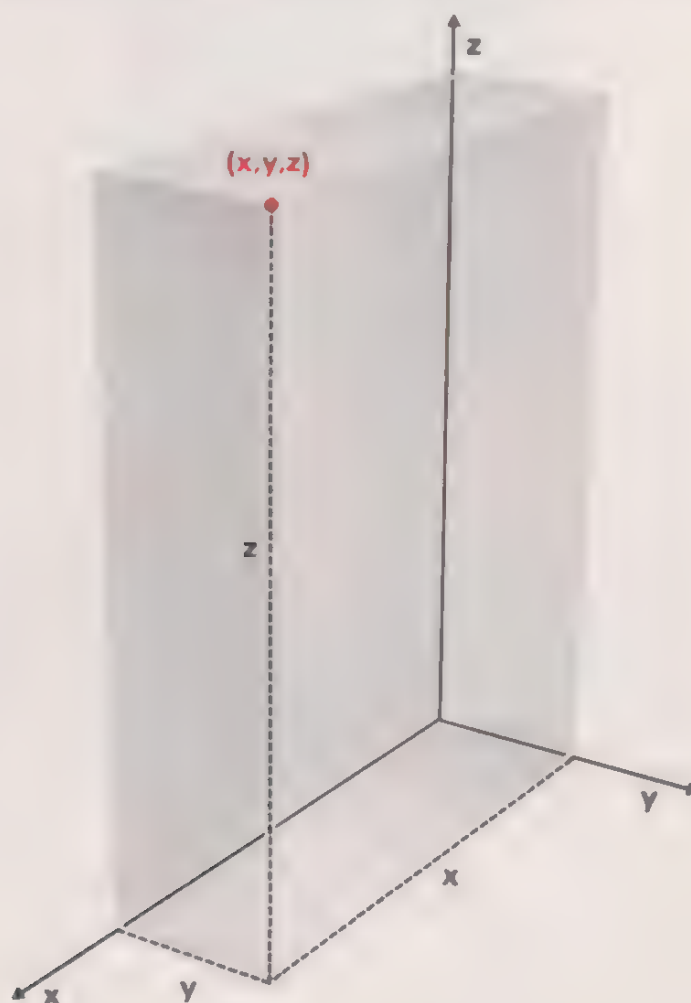
For example, we reach the point (3, 1, 5) if we start at the origin and proceed

>   3 units along the x-axis,
>   1 unit parallel to the y-axis,

and

>   5 units parallel to the z-axis.
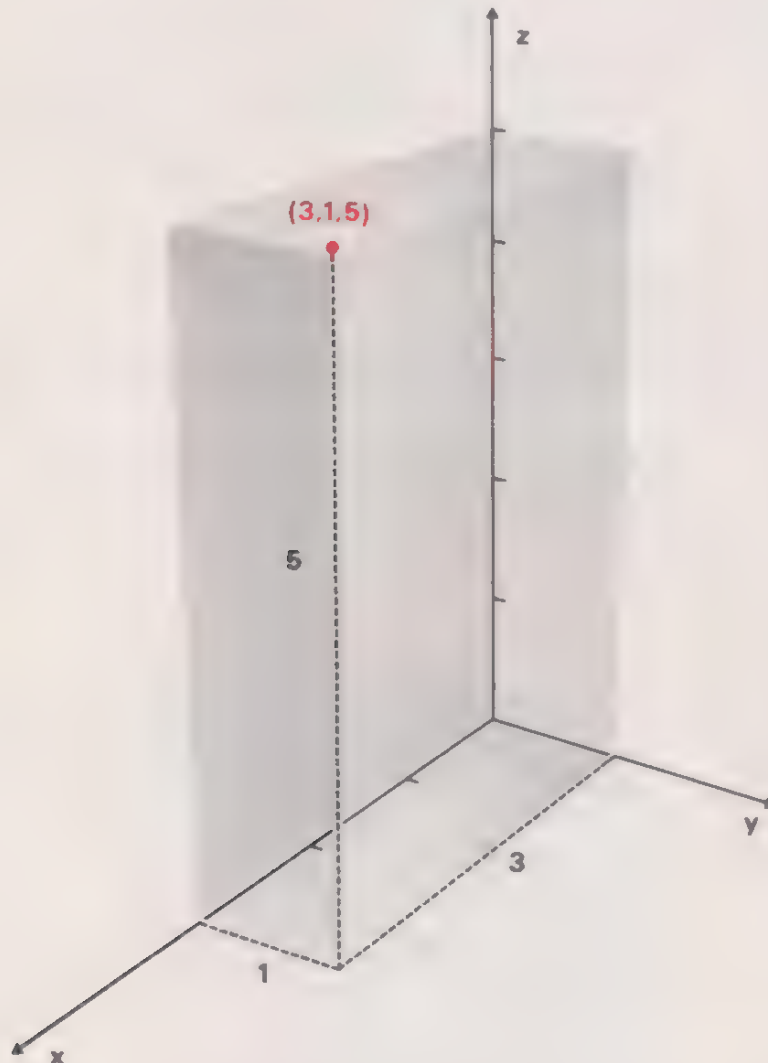
For functions of one variable, we know that a function gives rise to a graph (in the sense of a list), and this can be illustrated as a graph (in the sense of a picture). We now try to do something similar for a function of two variables. Let us look at a particular example of a function of two variables, and see how it gives rise first to a list and then to a picture.

Example 1

Consider the function

$$F : (x, y) \longmapsto \sqrt{x^2 + y^2} \qquad ((x, y) \in R \times R).$$

The ordered pair $(3, 4)$ is mapped to $\sqrt{3^2 + 4^2} = 5$, and this corresponds to the ordered pair $((3, 4), 5)$. (Notice that the first element of this pair is also a pair.) Similarly, the pair $(5, 12)$ maps to 13, and this corresponds to the pair $((5, 12), 13)$. If we put $F(x, y) = z$, then (x, y) maps to z, which gives rise to the pair $((x, y), z)$. In this way we can build up a table:

(x, y)	z
(3, 4)	5
(5, 12)	13
. . .	. . .

With the pair $((3, 4), 5)$ we can associate the point with co-ordinates $(3, 4, 5)$; with the pair $((5, 12), 13)$ we can associate the point with co-ordinates $(5, 12, 13)$; and so on. In this way, the function defines a set of ordered triples. Alternatively, we can think of the equation $z = F(x, y)$ as defining a *restriction* on the variables x, y and z. This restriction corresponds to a subset of $R \times R \times R$ (the set of all ordered triples of real numbers), namely the subset $\{(x, y, z) : z = F(x, y)\}$.

The surface corresponding to this function F is particularly easy to visualize, for in this case

$$z = F(x, y) = \sqrt{x^2 + y^2}$$
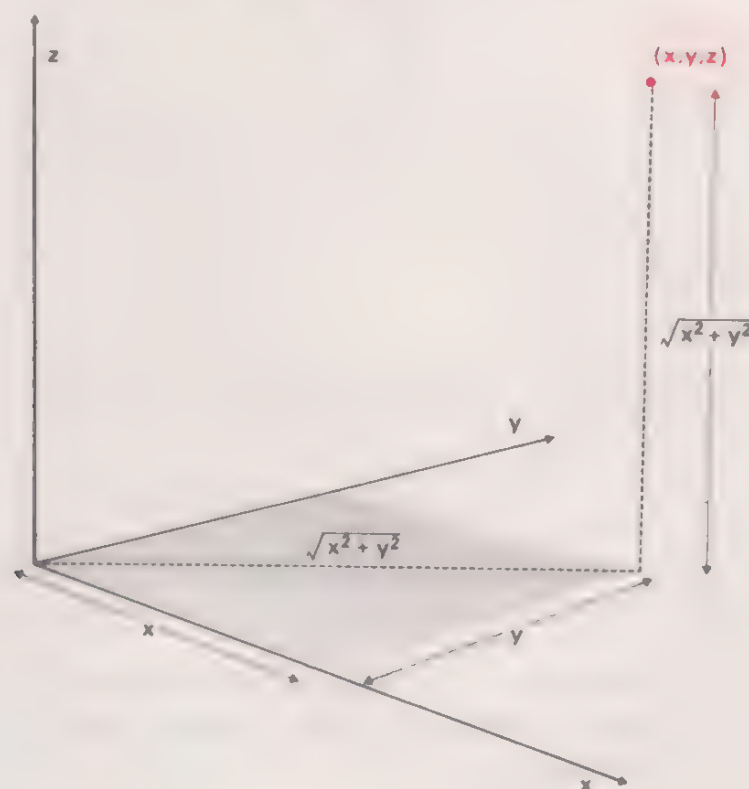
If we fix the value of z (corresponding to the vertical height in the diagram) and look at all the points at this height whose co-ordinates satisfy

$$z = \sqrt{x^2 + y^2},$$

then we find that they are all the same distance, $\sqrt{x^2 + y^2}$, from the z-axis; that is, they lie on a circle.

We can describe the surface in words by saying that we move from the origin in any horizontal direction, then vertically through the same distance to reach the surface. This surface is a cone with its vertex at the origin.
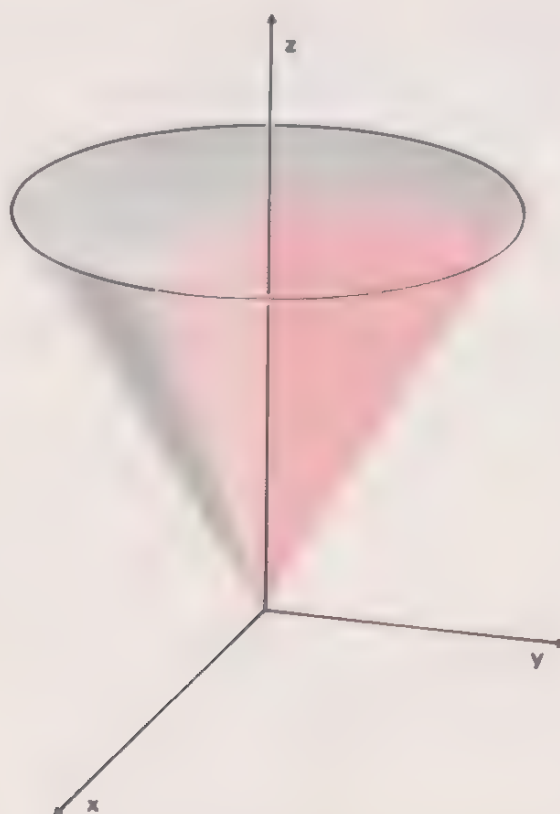
In general, let F be any function of two (real) variables:

$$F : (x, y) \longmapsto z \qquad ((x, y) \in R \times R).$$

Then to each ordered pair (a, b) we can associate the point Q with co-ordinates $(a, b, F(a, b))$.

In general, then, a function F of two variables defines a subset of $R \times R \times R$ which is often a surface. It is this surface which is the generalization of the pictorial graph of a function of one real variable.
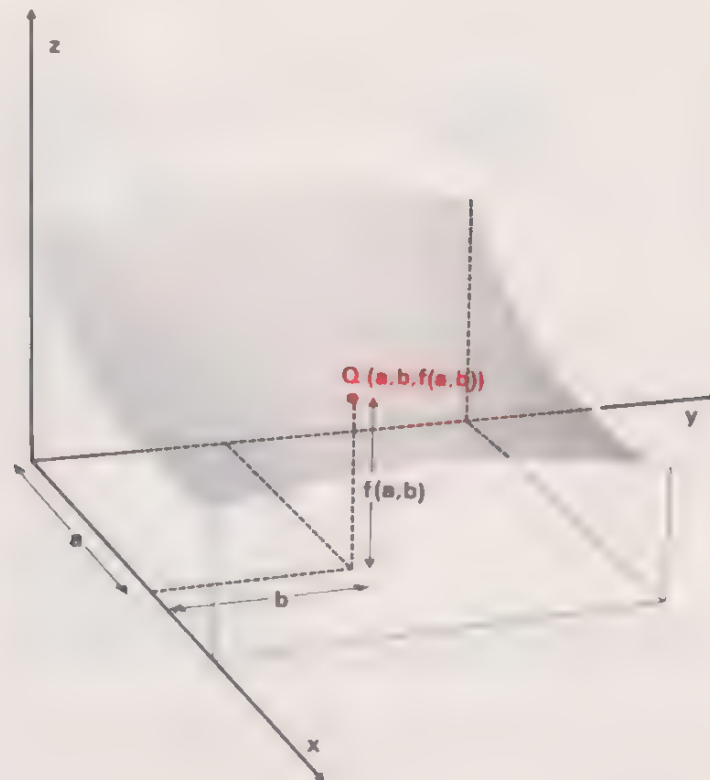
Exercise 1

Indicate on a diagram the sets of points with co-ordinates (x, y, z) satisfying:

  (i) $x = 0$,
 (ii) $y = 0$,
(iii) $x = y = 0$.

Exercise 2

Mark on a diagram the set of points with co-ordinates (x, y, z) in $R \times R \times R$ which corresponds to the condition:

$$2x - y = 0.$$

Before we find the equation which defines a general plane (which we do in the next section), we would like to give you some reasons for our interest in the subject. At first sight the following examples have nothing to do with planes, but a closer examination will reveal the connection.

Contour Lines

It is difficult to give an impression of a three-dimensional object on a two-dimensional piece of paper. One way of overcoming this difficulty, which is shown in the following examples, may be quite familiar to you, and we can develop it into a useful tool for examining functions of two variables.

Example 2   Barometric Pressure

ATLANTIC WEATHER MAP

SITUATION AT NOON JAN.22
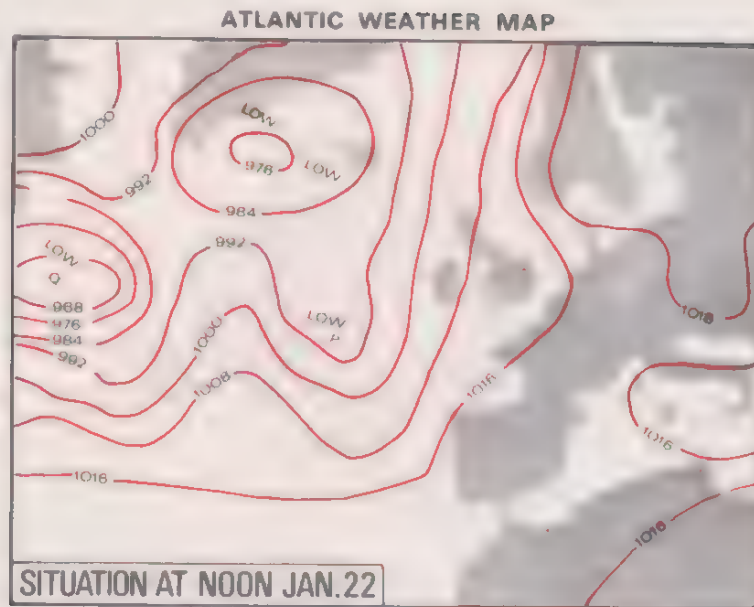
The red curves on the weather map join the points of equal barometric pressure. (The pressure varies with height, but the values shown refer to the pressure at sea-level.) The function illustrated in this case is

$$P : \text{(point on the map)} \longmapsto \text{(barometric pressure at the corresponding point on the earth's surface)}.$$

Example 3   Ordnance Survey Maps
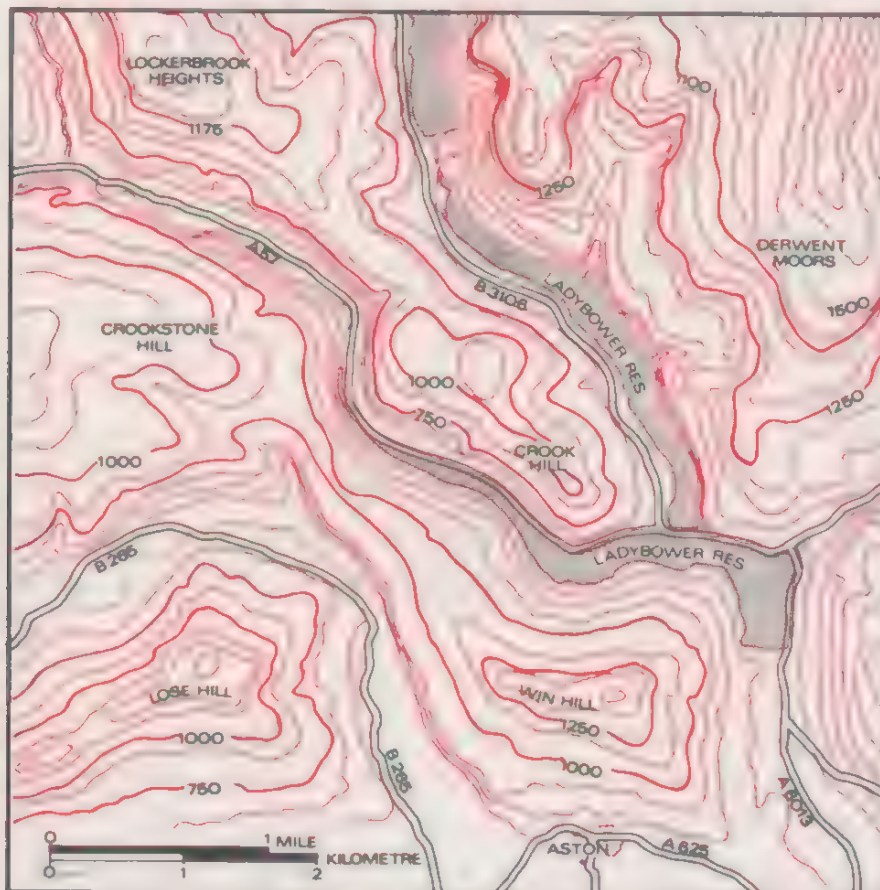
The cartographer has only a flat piece of paper, but he does his best to give an impression of the shape of the land surface by showing us the contour lines; in other words, he joins the points of equal height above sea level.

The function which is illustrated in this case is

$$h : (\text{point on the map}) \longmapsto (\text{height of the corresponding point above sea level})$$

Example 4

Suppose that we take a circular cylinder of unit radius which has its axis vertical (along the z-axis) and intersect it with a horizontal plane 5 units above the xy-plane, as shown in the following diagram.

The two surfaces (the cylinder and the plane) meet in a curve, which is in fact a horizontal circle 5 units above the xy-plane.
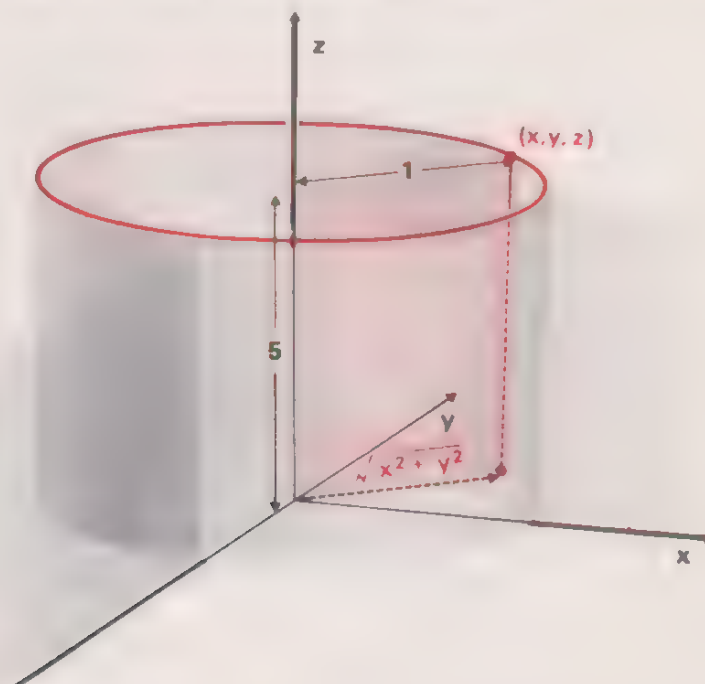
Any point on the cylinder is one unit from the z-axis and therefore $\sqrt{x^2 + y^2} = 1$. The equation of the cylinder is therefore $\sqrt{x^2 + y^2} = 1$, by which we mean that the set of all points with co-ordinates (x, y, z) in $R \times R \times R = R^3$, satisfying this equation, lie on the cylinder.

The equation of the plane is $z = 5$, and the two equations taken together:

$$\sqrt{x^2 + y^2} = 1$$

$$z = 5$$

determine the set of points lying on the red circle.

Another way of writing this is as follows: denote a point P with co-ordinates (x, y, z) by $P(x, y, z)$: then the cylinder is the set

$$A = \{P(x, y, z):(x, y, z)\in R^3, \sqrt{x^2 + y^2} = 1\};$$

the plane is the set

$$B = \{P(x, y, z):(x, y, z)\in R^3, z = 5\};$$

and the circle is $A \cap B$.

You may say that our two equations imply that $x^2 + y^2 = \dfrac{z}{5}$, so isn't this the equation of the circle? If so, compare the following two sets in $R \times R \times R$:

(i) $A \cap B$, the set of points with co-ordinates (x, y, z) which satisfy

$$\sqrt{x^2 + y^2} = 1$$

and

$$z = 5:$$

(ii) C, the set of points with co-ordinates (x, y, z) which satisfy

$$x^2 + y^2 = \frac{z}{5}.$$

Notice that every point in $A \cap B$ has $z = 5$, so that it must have co-ordinates of the form $(x, y, 5)$, whereas the point with co-ordinates $(2, 2, 40)$, for instance, lies in C but not in $A \cap B$. Every point in $A \cap B$ satisfies the conditions for C, but the converse is not true, so that $A \cap B$ is a proper subset of C. This is to be expected because $A \cap B$ implies a stronger restriction in $R \times R \times R$ than does C. To belong to C, (x, y, z) must be such that $x^2 + y^2$ is the same as $\dfrac{z}{5}$. To belong to $A \cap B$, (x, y, z) must

be such that, not only is $x^2 + y^2$ the same as $\dfrac{z}{5}$, but also both expressions have the value 1. In fact the points in C lie on what is called a paraboloid of revolution, and this paraboloid contains the red circle.
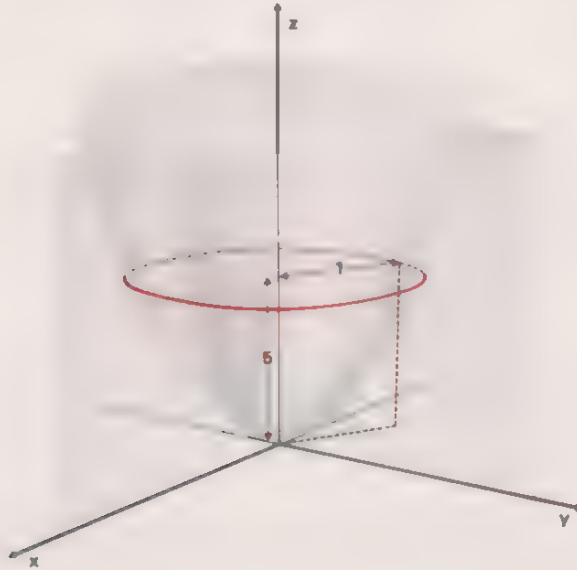
A Generalization

Suppose that we are given an arbitrary function of two variables, F, with domain the xy-plane, and we intersect the surface

$$z = F(x, y)$$

with the horizontal plane

$$z = c$$
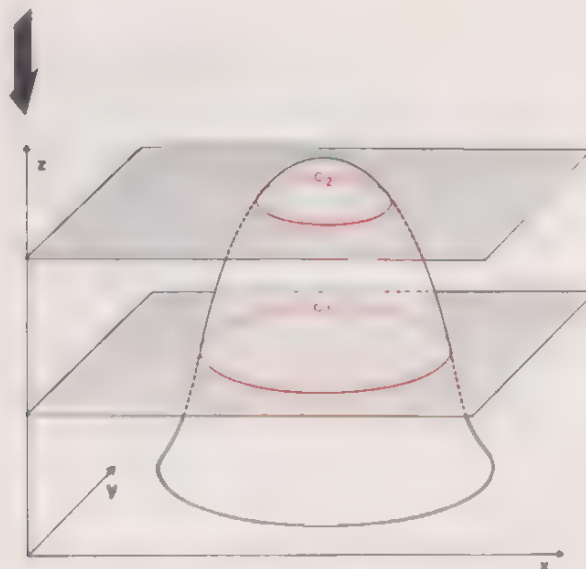
Various contour lines shown in red for various values of c.

The resulting curve is called the contour line corresponding to the height c. Taking various values of c will give a set of contour lines, which, when viewed from above (looking down the z-axis), could look like this:
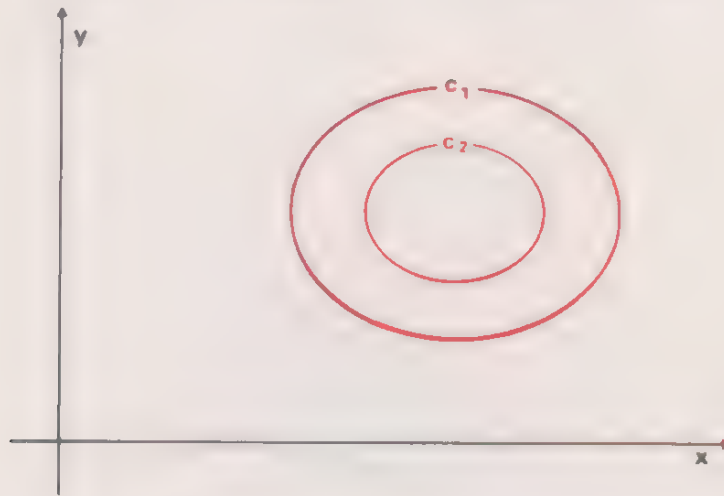
This is simply a general version of the pressure and altitude diagrams which we used to introduce these ideas.

The previous examples have shown how planes parallel to the xy-plane (horizontal planes) can be used to describe surfaces, but we intend to use planes parallel to the z-axis (vertical planes) too.

Consider the intersection of our arbitrary surface defined by

$$z = F(x, y)$$

with the plane

$$y = b.$$
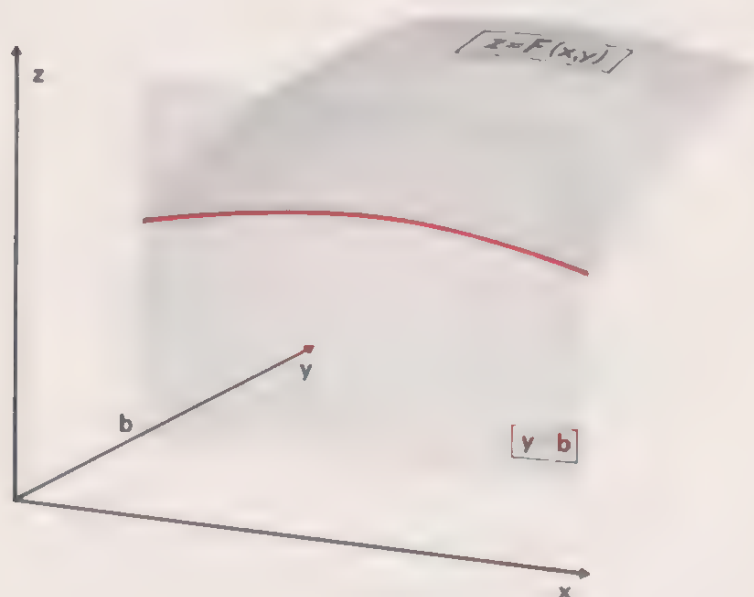
The effect is rather like slicing a Dutch cheese: for each slice the red rind of the cheese forms a different curve, and for each value of b in the above diagram we get a new red curve. The advantage of this idea is that it reduces a surface, which is difficult to draw, to a set of curves, each lying in a plane, which we *can* draw on a piece of paper. Mathematically speaking, we have reduced a function of two variables to a whole set of functions of one variable each corresponding to a particular value of b, and a particular red curve.

2.2   The General Equation of a Plane

Up to this point we have been finding the surfaces corresponding to given equations and functions, but now we want to put the problem in reverse. Can we find equations of given surfaces? What is the equation of a plane? This is an essential step on our way to solving optimization problems for functions of two variables.

Suppose that the plane passes through the point (a, b, c) and that it is inclined at an angle A in the x direction and an angle B in the y direction.
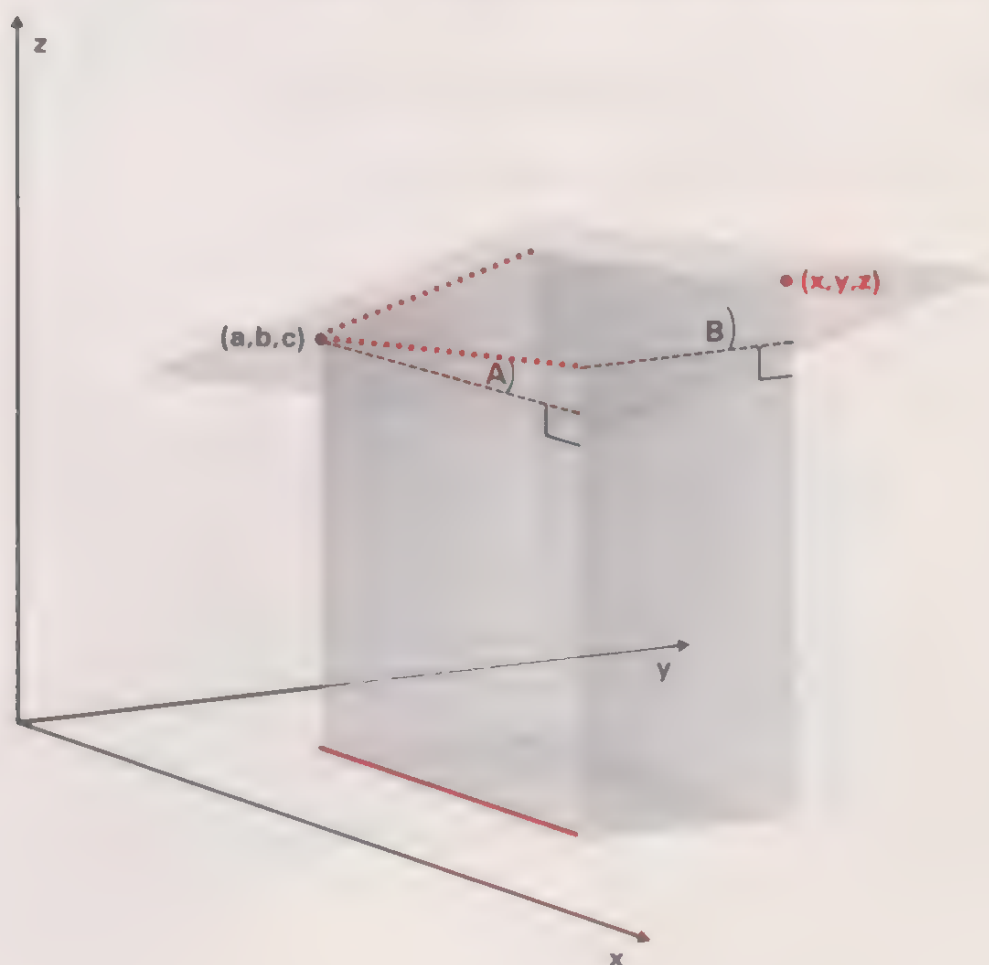
To help us obtain the expression for z in terms of x and y, we give below a diagram which shows the two "nearer" sides of the "box" in the figure above opened out and flattened into a plane.
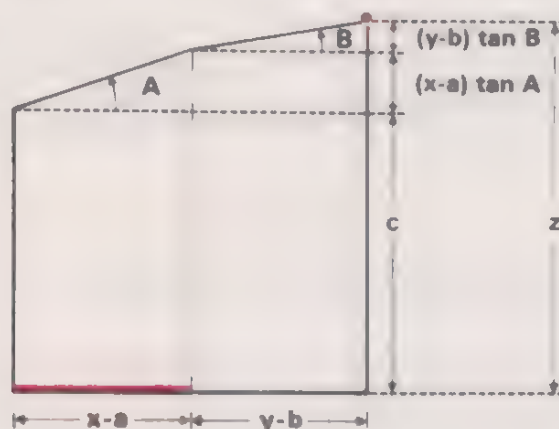
The value of z corresponding to an arbitrary choice of (x, y) is simply the result of adding the three terms on the right of the above diagram,

$$z = c + (x - a)\tan A + (y - b)\tan B, \qquad \text{Equation (1)}$$

which is the required equation of the plane if neither A nor B is $\dfrac{\pi}{2}$.

If either A or B is $\dfrac{\pi}{2}$, then the plane is perpendicular to the xy-plane, and looking down the z-axis we see only the line PQ.
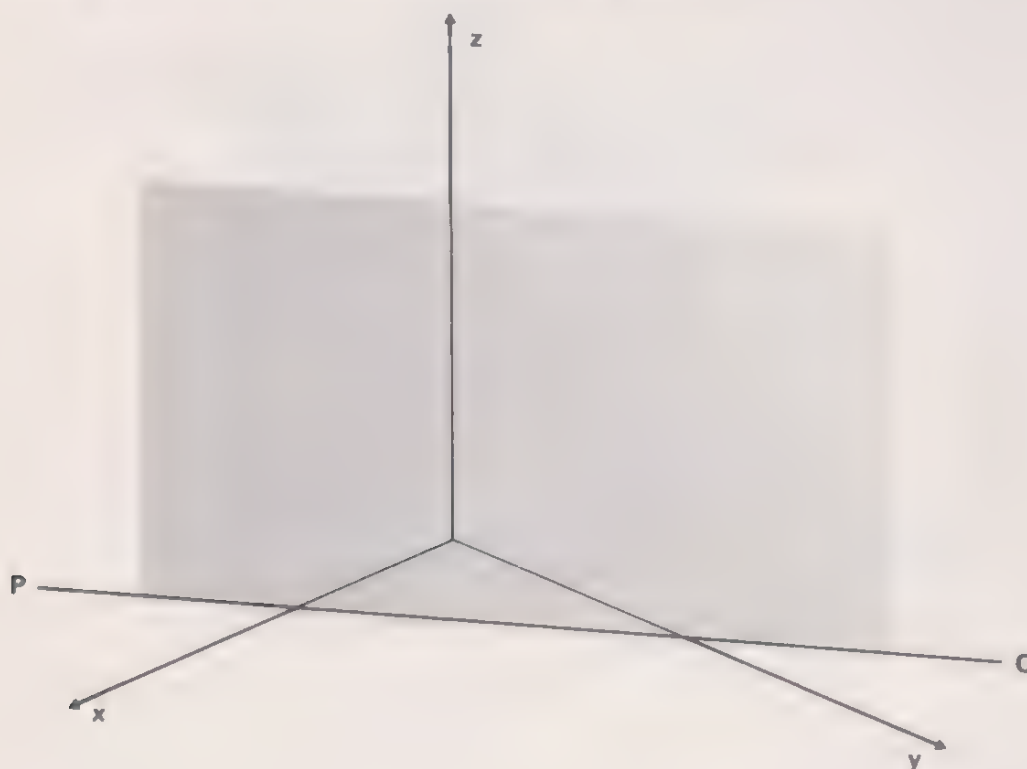
In the xy-plane we could represent the line PQ by the equation

$$\alpha x + \beta y + \delta = 0.$$

In $R \times R \times R$ this equation represents the plane.

Equation (1) and Equation (2) are particular cases of the equation

$$\alpha x + \beta y + \gamma z + \delta = 0$$

where α, β, γ and δ are real numbers independent of x, y, z (see also the following exercise). This is the general equation of a plane. Notice particularly that the plane is horizontal (that is, parallel to the xy-plane) if $\alpha = \beta = 0$ (in other words when the angles A and B are both zero).

Exercise 1

(i) What values should we take for α, β, γ and δ in order to make Equation (3) identical to Equation (1)?

(ii) At what points does the plane $\alpha x + \beta y + \gamma z + \delta = 0$ meet each of the three co-ordinate axes?

(iii) The equation $\lambda(x - a) + \mu(y - b) = 0$ represents a plane perpendicular to the xy-plane which passes through the points (a, b, z) for any value of z. What effect does it have on the plane if we vary the values of λ and μ?

2.3 Partial Derivatives

We have derived the general equation of a plane, but we really need the equation of the *tangent plane* at a point on a given surface. We can then imagine this plane moving over the surface, and we hope that this notion will give us a technique for finding the maximum (or minimum) value of the corresponding function, just as a similar idea helped for functions of one variable. For the moment we need something like the derivative of a function of one variable, which was useful when discussing rate of change. The corresponding concept which we are going to examine is that of a *partial derivative*.

First let us give an intuitive idea of the concept of partial derivative. Imagine yourself standing at a crossroads on a hillside, the roads running East–West and North–South. Roughly speaking, the slopes of the East–West road and the North–South road are the partial derivatives of the function, represented by the hillside, at the point where the roads cross. If the crossroads happened to be at the top of a hill then each of the slopes

would be zero. It is this intuitive idea that we want to make precise, and the following example will lead us in the right direction.

Example 1

Consider the surface representing the function

$$F:(x, y) \longmapsto \sqrt{1 - (x^2 + y^2)} \qquad ((x, y) \in R \times R, x^2 + y^2 \leqslant 1).$$

<div align="right">**Equation (1)**</div>

The domain of F is represented in the xy-plane by the points on and within the circle with radius 1, centred at the origin.

If we let $z = \sqrt{1 - (x^2 + y^2)}$, then it follows that $x^2 + y^2 + z^2 = 1$.
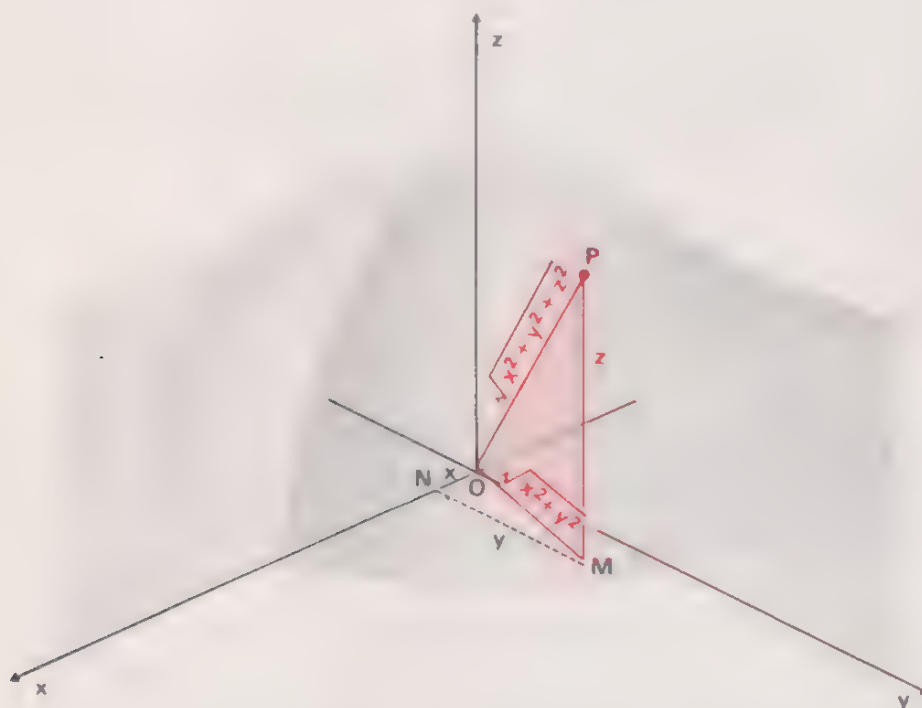
The distance of any point, $P(x, y, z)$, in $R \times R \times R$ from the origin is $\sqrt{x^2 + y^2 + z^2}$. This can be seen in the diagram, first by using Pythagoras's Theorem in the triangle ONM, and then in the triangle OMP. Since points on the surface satisfy the equation $x^2 + y^2 + z^2 = 1$, it follows that any point P lying on the surface must be at unit distance from the origin, and since z is always positive, Equation (1) represents a hemisphere.

In terms of our intuitive discussion, this hemisphere is the hillside. We are now going to choose a point Q on this hillside and assume that Q is our crossroads, with the roads through Q in the vertical planes through Q parallel to the x and y axes. Any point Q will do, and to illustrate the idea we choose Q to be the point with co-ordinates $\left(\dfrac{1}{2}, 0, \dfrac{\sqrt{3}}{2}\right)$.
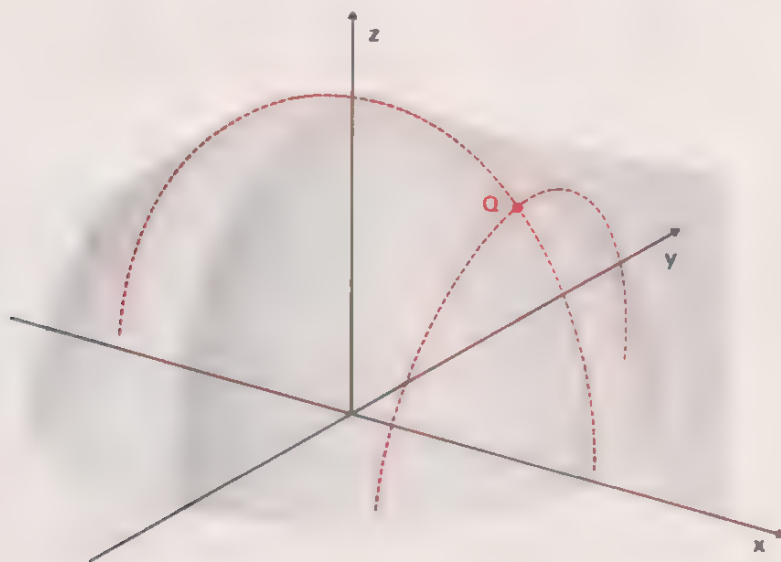
The point Q with co-ordinates $\left(\dfrac{1}{2}, 0, \dfrac{\sqrt{3}}{2}\right)$ lies on the surface and on the plane $y = 0$. If we were to cut the hemisphere with the plane $y = 0$ through Q, and then look along the y-axis, we would see the semi-circle shown in red in the following diagram:
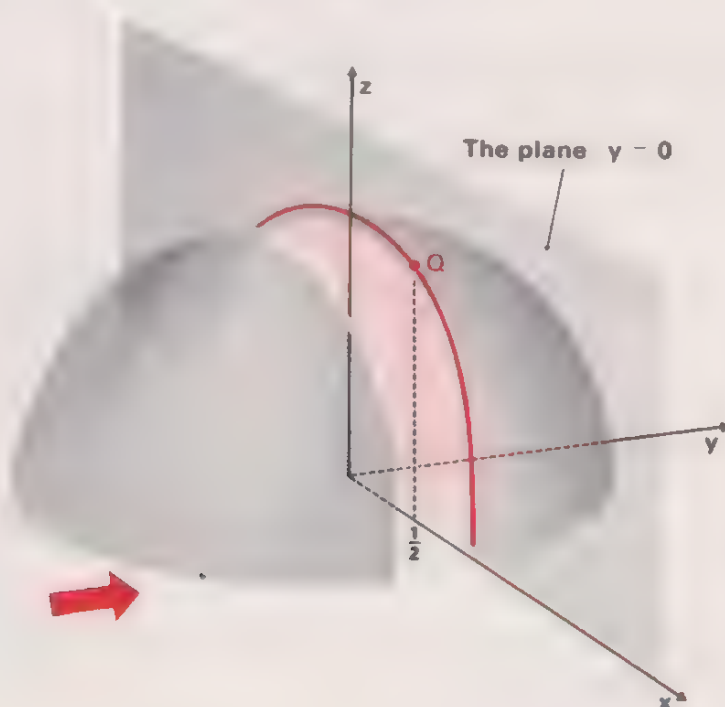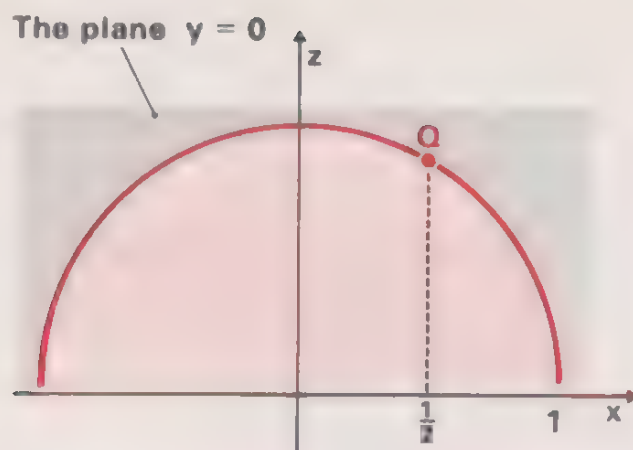
You can imagine the red curve to be the road through Q in the x-direction; later we shall find it very useful to be able to calculate the slope of such curves at an arbitrary point Q.

Next we find the slope of the red semi-circle at the point Q with co-ordinates $\left(\dfrac{1}{2}, 0, \dfrac{\sqrt{3}}{2}\right)$. The semi-circle is determined by the equations:

$$z = \sqrt{1 - (x^2 + y^2)}$$

$$y = 0$$

so that on the curve we have:

$$z = \sqrt{1 - x^2} \qquad (x \in [-1, +1]).$$

But a relationship like this defines a function, f_1 say:

$$f_1 : x \longmapsto \sqrt{1 - x^2} \qquad (x \in [-1, +1]).$$

Since we are interested in the slope at Q, we calculate the derivative, $f_1'(x)$. We get

$$f_1'(x) = \frac{-x}{\sqrt{1 - x^2}}$$

which takes the value $-\dfrac{1}{\sqrt{3}}$ when $x = \dfrac{1}{2}$.

Suppose now that we want the slope of the other road at Q. First intersect the hemisphere with the plane $x = \frac{1}{2}$:
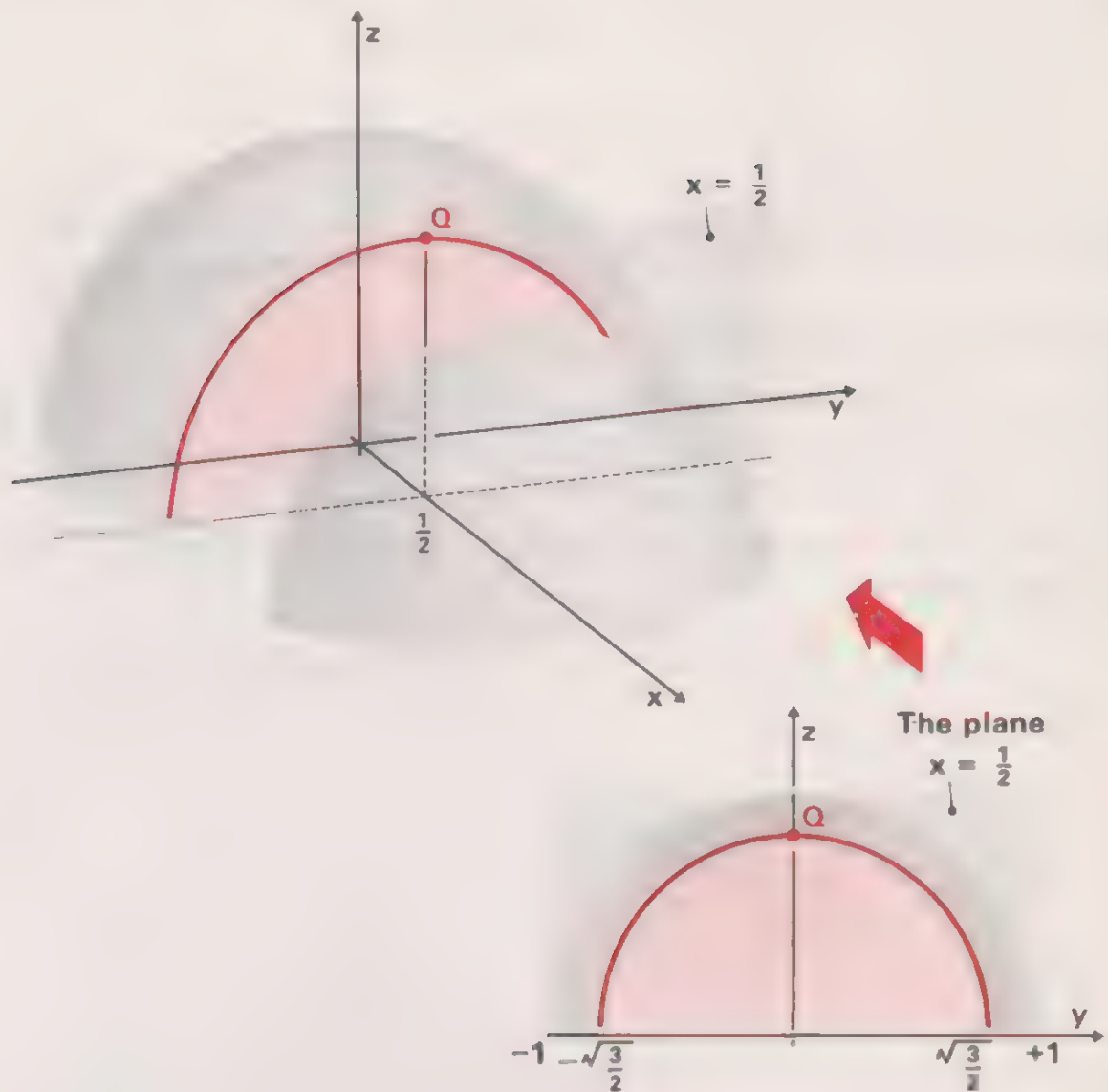
The red curve in the diagram is determined by the equations:

$$z = \sqrt{1 - (x^2 + y^2)}$$

$$x = \tfrac{1}{2}$$

so that on the curve we have:

$$z = \sqrt{1 - (\tfrac{1}{4} + y^2)}$$

$$= \sqrt{\tfrac{3}{4} - y^2} \qquad \left(y \in \left[-\frac{\sqrt{3}}{2}, +\frac{\sqrt{3}}{2} \right] \right).$$

This relationship defines a function, f_2 say:

$$f_2 : y \longmapsto \sqrt{\tfrac{3}{4} - y^2} \qquad \left(y \in \left[-\frac{\sqrt{3}}{2}, +\frac{\sqrt{3}}{2} \right] \right). \qquad \cdot$$

Again we calculate the derivative, $f'_2(y)$. We get

$$f'_2(y) = \frac{-y}{\sqrt{\frac{3}{4} - y^2}}$$

which takes the value 0 when $y = 0$ (the slope of the curve at Q seemed likely to be zero from the diagram, so this should be no surprise).

Definition of Partial Derivatives

Having worked through the previous example, you may well feel that there must be a quicker way, and indeed there is. It is, however, the intuitive ideas of that example which point the way.

Suppose that we examine the function F from a slightly different point of view. We know that

$$F:(x, y) \longmapsto \sqrt{1 - (x^2 + y^2)} \qquad ((x, y) \in \{(x, y): x^2 + y^2 \leqslant 1\}).$$
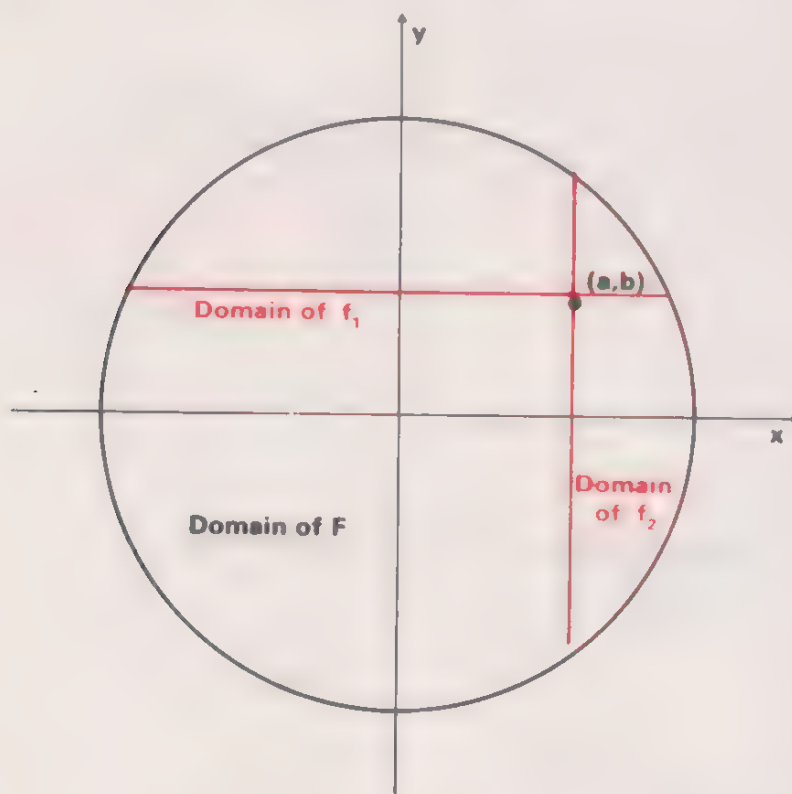
If we keep y constant, $y = b$ say, then we obtain a new function (of one variable):

$$f_1:x \longmapsto \sqrt{1 - (x^2 + b^2)} \qquad (x \in [-\sqrt{1 - b^2}, +\sqrt{1 - b^2}]),$$

and

$$f'_1(x) = \frac{-x}{\sqrt{1 - (x^2 + b^2)}}.$$

Similarly, if we keep x constant, $x = a$ say, then we obtain a new function (of one variable):

$$f_2 : y \longmapsto \sqrt{1 - (a^2 + y^2)} \qquad (y \in [-\sqrt{1 - a^2}, + \sqrt{1 - a^2}]),$$

and

$$f_2'(y) = \frac{-y}{\sqrt{1 - (a^2 + y^2)}}.$$

The slopes of the roads through the point Q, $\left(\dfrac{1}{2}, 0, \dfrac{\sqrt{3}}{2}\right)$, running parallel to the x and y axes are given by the derivatives $f_1'(\tfrac{1}{2})$ and $f_2'(0)$ respectively (taking $a = \tfrac{1}{2}$ and $b = 0$).

The expression $f_1'(x)$ gives the slope of the surface (defined by the function F) in the direction of the x-axis at the point (x, b); that is, $f_1'(x)$ is the rate of change of F with respect to x, when y has the constant value b.

Similarly, $f_2'(y)$ gives the slope of the surface in the direction of the y-axis at the point (a, y); that is, $f_2'(y)$ is the rate of change of F with respect to y, when x has the constant value a.

We chose to consider the point Q with co-ordinates $\left(\dfrac{1}{2}, 0, \dfrac{\sqrt{3}}{2}\right)$, so we took $a = \tfrac{1}{2}$ and $b = 0$. We would like to know the corresponding slopes (rates of change) at *any* point on the surface defined by F; that is, we now wish to *vary* a and b. This means that we need to express the slopes in terms of functions of *two* variables. So we define two *new* functions, F_1' and F_2', by the equations:

$$F_1'(x, y) = \frac{-x}{\sqrt{1 - (x^2 + y^2)}} \qquad ((x, y) \in \{(x, y) : x^2 + y^2 \leqslant 1\})$$

and

$$F_2'(x, y) = \frac{-y}{\sqrt{1 - (x^2 + y^2)}} \qquad ((x, y) \in \{(x, y) : x^2 + y^2 \leqslant 1\}).$$

We are thus led to the following *definition of partial derivatives* of a function, F, of two variables x and y.

The partial derivative of F with respect to the first variable, x, at (x, y) is

$$F'_1(x, y) = \lim_{h \to 0} \frac{F(x + h, y) - F(x, y)}{h}$$

The partial derivative of F with respect to the second variable, y, at (x, y) is

$$F'_2(x, y) = \lim_{k \to 0} \frac{F(x, y + k) - F(x, y)}{k}$$

In order to find the two partial derivatives, we simply keep each of the variables fixed in turn and differentiate with respect to the other.

Example 2

If

$$G : (x, y) \longmapsto 2xy + x^2 \qquad ((x, y) \in R \times R),$$

then, differentiating with respect to x at (x, y), we regard y as constant and get

$$G'_1(x, y) = 2y + 2x,$$

and differentiating with respect to y at (x, y), we regard x as constant and get

$$G'_2(x, y) = 2x.$$

Working directly from the definitions, we have :

$$G'_1(x, y) = \lim_{h \to 0} \left(\frac{2y(x + h) + (x + h)^2 - (2xy + x^2)}{h} \right)$$

$$= \lim_{h \to 0} \left(\frac{2yh + 2xh + h^2}{h} \right)$$

$$= 2y + 2x,$$

and

$$G'_2(x, y) = \lim_{k \to 0} \left(\frac{2x(y + k) + x^2 - (2xy + x^2)}{k} \right)$$

$$= \lim_{k \to 0} \left(\frac{2xk}{k} \right)$$

$$= 2x$$

Exercise 1

Find the partial derivatives at (x, y) of the functions defined by the following equations; each function has domain $R \times R$.

(i) $F(x, y) = x^2 + y^2$
(ii) $G(x, y) = x \exp(xy)$

Alternative Notation

There are various notations for the partial derivatives; the most common is $\dfrac{\partial F}{\partial x}$ for what we write as $F'_1(x, y)$. This notation arose presumably because of the commonly used notation $\dfrac{df}{dx}$ for the derivative of a function, f, of one variable. If you use the $\dfrac{\partial F}{\partial x}$ notation, then you must be extremely careful later. For example, it is not generally true that $\dfrac{\partial F}{\partial x} \cdot \dfrac{\partial x}{\partial t}$ is simply $\dfrac{\partial F}{\partial t}$, as the notation would suggest.

The alternative notation F_x or F_y is also very common.

2.4 The Tangent Plane

You probably have an intuitive notion of what we mean by the *tangent plane* at a particular point on a surface. It is, after all, the plane which sits comfortably on the surface at the point in question. Once again, we assume that our surfaces are smooth with no sharp projections. It would, for example, be difficult to decide where the tangent plane should be on the apex of a church steeple.

On the other hand, it is quite easy to imagine a tangent plane at a point on a smooth sphere.

We shall now define the tangent plane at any point on a smooth surface. Suppose that we are given a surface defined by:

$$F : (x, y) \longmapsto F(x, y) \qquad ((x, y) \in R \times R),$$

and we wish to define the tangent plane at the point Q with co-ordinates

$(a, b, F(a, b))$. We have seen in Exercise 2.2.1 (iii) that the equation

$$\lambda(x - a) + \mu(y - b) = 0$$

defines a plane which passes through Q and is perpendicular to the xy-plane.
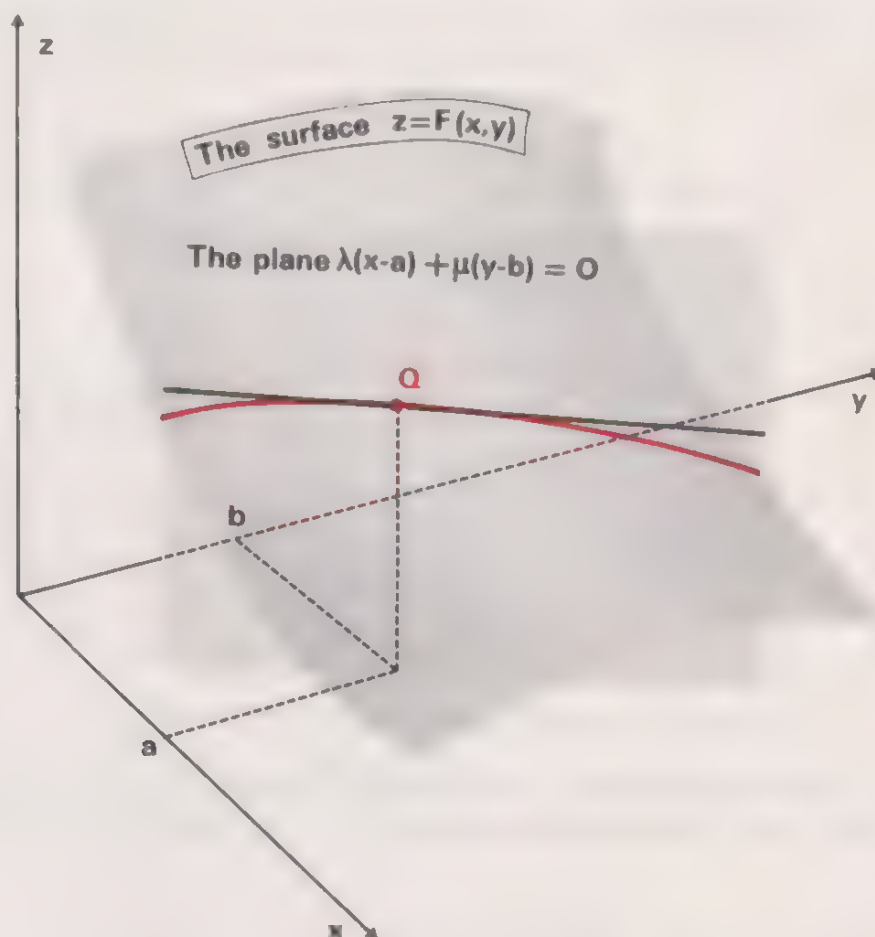
The intersection of this plane with the surface will be a curve (shown in red on the diagram). This curve passes through Q and has a tangent line (shown by a heavy black line) at Q. If we vary the values of λ and μ, the plane will rotate about the vertical line through Q, and each pair of values of λ and μ will give us such a tangent line. If *all* these tangent lines at Q lie in a plane, then we call this plane the tangent plane at Q.

The Equation of the Tangent Plane

Our assumption that the surfaces we meet are smooth is intended to imply that there *is* a tangent plane to the surface $z = F(x, y)$ at Q, but how can we find its equation?

Suppose that we take the particular values $\mu = 1$, $\lambda = 0$, in the equation $\lambda(x - a) + \mu(y - b) = 0$.

Then we simply get the equation

$$y = b,$$

and the slope of the corresponding curve of intersection at Q is $F_1'(a, b)$. In other words, $F_1'(a, b)$ is the slope of the tangent to this curve at Q.

The curve of intersection is shown in red on the following diagram, and the tangent to this curve at Q is shown by a heavy black line.
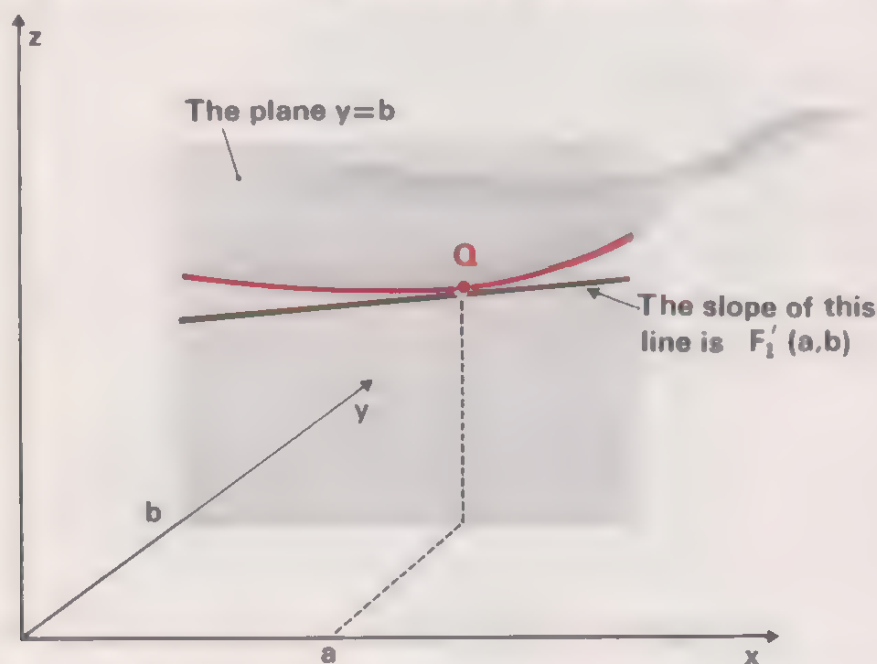
The slope of the tangent line in the above diagram is simply $F_1'(a, b)$, and similarly the slope of the corresponding line in the y-direction is $F_2'(a, b)$.
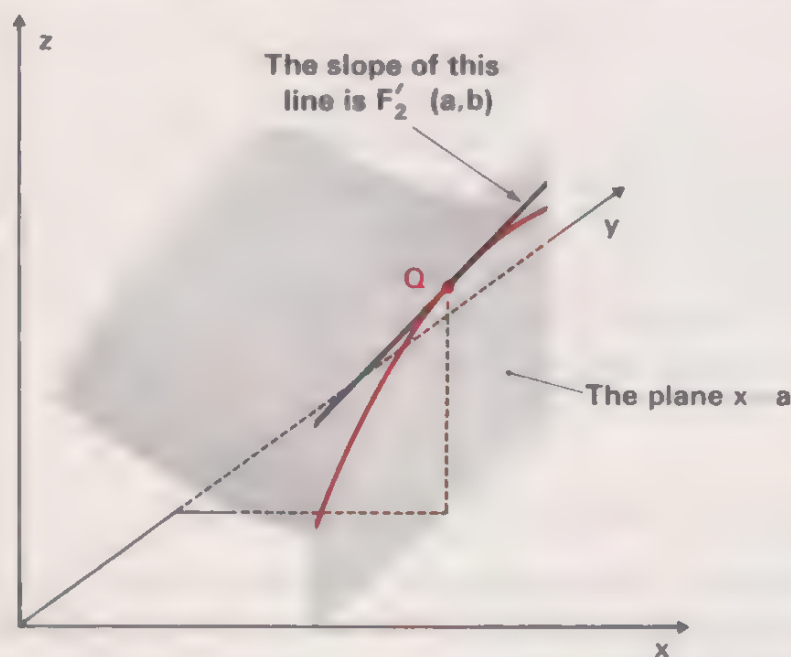
We can now return to the figure on page 37 which we show below (from a different viewpoint). This time we use it to find the equation of the tangent plane at Q.
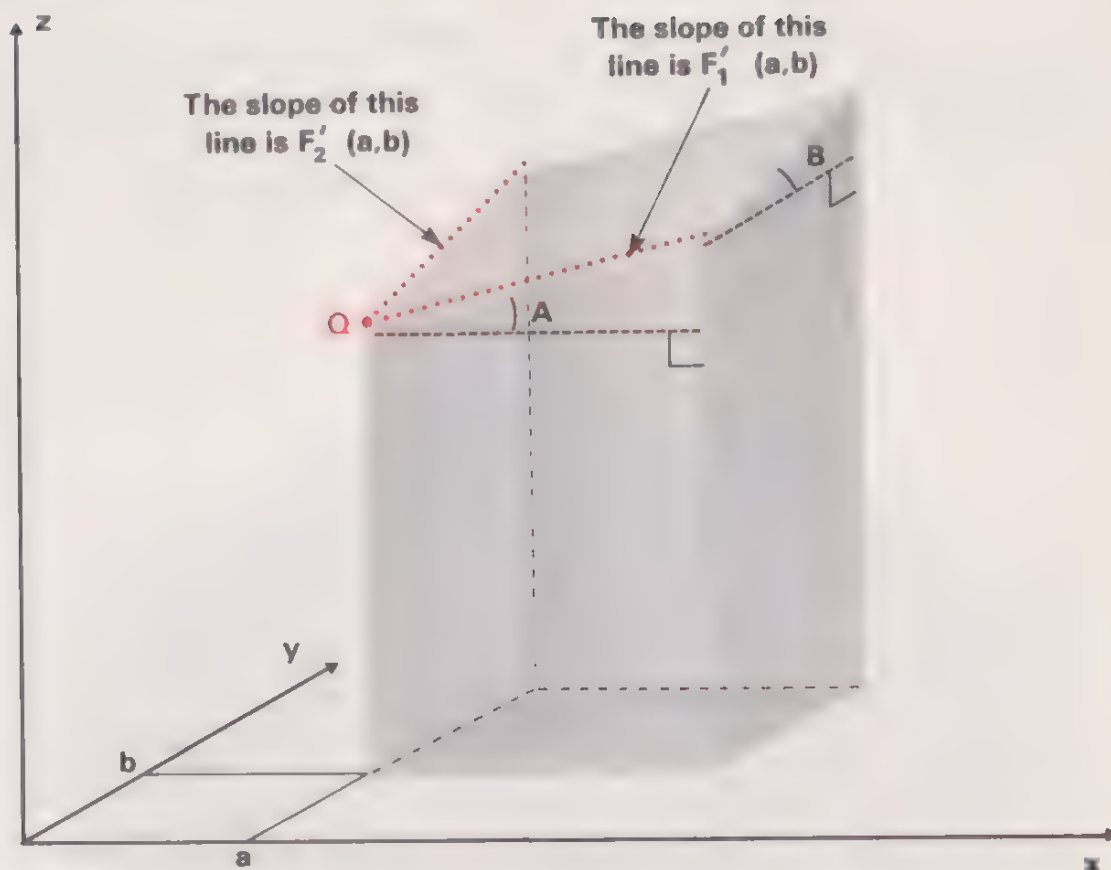
In the formula on page 38, we simply put $\tan A = F_1'(a, b)$, $\tan B = F_2'(a, b)$ and $c = F(a, b)$. We can see from the last diagram that any point on the tangent plane has co-ordinates (x, y, z) which satisfy the equation

$$z = F(a, b) + F_1'(a, b)(x - a) + F_2'(a, b)(y - b)$$

and this is the equation of the tangent plane to the surface at $(a, b, F(a, b))$.

Exercise 1

For each of the following functions, find the equation of the tangent plane at the point on the surface corresponding to the pair (a, b) (each function has domain $R \times R$).

(i) $F : (x, y) \longmapsto x^2 + y^2$

(ii) $G : (x, y) \longmapsto x \exp(xy)$

(Use the results of Exercise 2.3.1.)

2.5 Optimizing Functions of Two Variables

There is no doubt that finding the overall maximum (or minimum) of a function of two variables is, in general, harder than finding the overall maximum (or minimum) of a function of one variable. Some might say, "more than twice as hard". For a differentiable function f of one variable

with domain $[a, b]$ we merely locate the stationary points, investigate
their nature, and then examine the values of $f(a)$ and $f(b)$. The domain of
a (real) function of two variables is a subset of $R \times R$, and instead of just
two end-points we are now likely to have a curve as the boundary of our
domain.
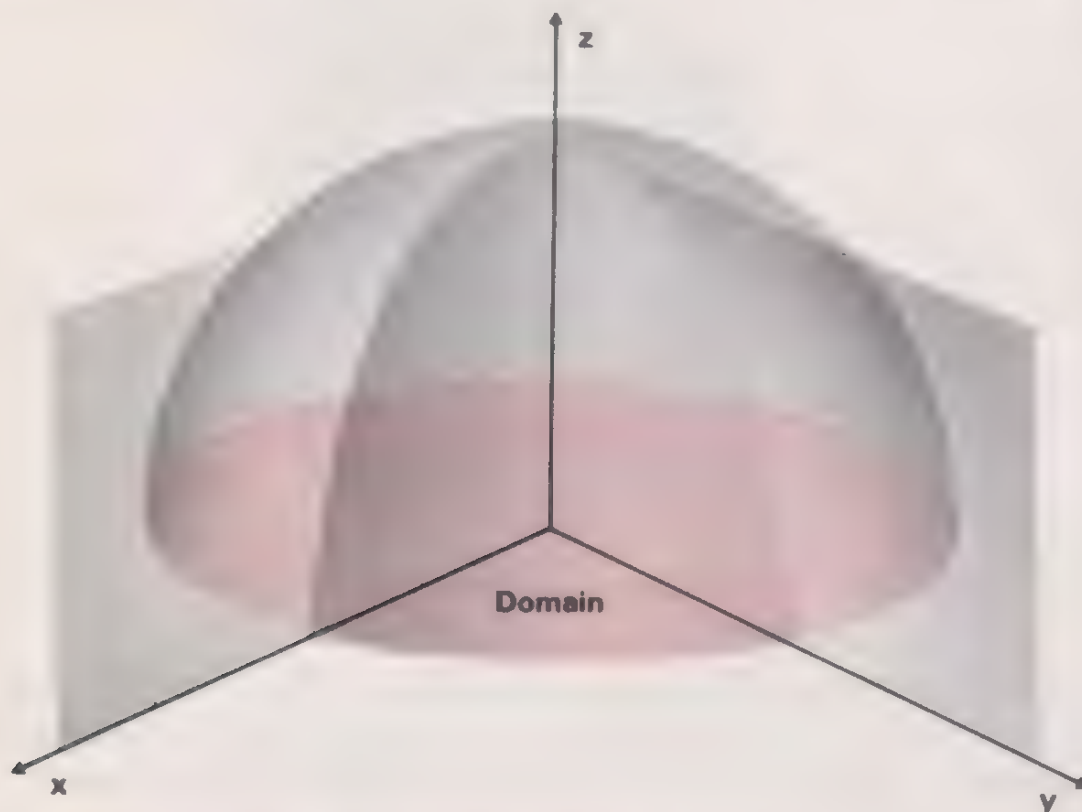
For example, we have already discussed the function

$$F : (x, y) \longmapsto \sqrt{1 - (x^2 + y^2)} \qquad ((x, y) \in \{(x, y) : x^2 + y^2 \leqslant 1\})$$

The domain has the circle $x^2 + y^2 = 1$ in the xy-plane as its boundary.

Suppose that we wish to find the overall maximum (or minimum) value
of the images of a function F with domain, A, a subset of $R \times R$. The points
where the tangent plane is parallel to the xy-plane, on the surface defined
by $z = F(x, y)$, are clearly going to be of interest. This leads us to our
next definition.

If $F'_1(a, b) = 0$ and $F'_2(a, b) = 0$, then (a, b) is called a stationary point of F.

Notice that since the tangent plane at a stationary point is parallel to the
xy-plane, its equation is simply $z = F(a, b)$.

Local Maxima and Minima

You may find the precise definitions of *local maximum* and *local minimum*
a little hard to digest, so we give intuitive definitions first.

If (a, b) is a point in the domain of F, and if $F(x, y) \leqslant F(a, b)$ for all (x, y)

in the domain of F sufficiently close to (a, b), then we say that F has a local maximum at (a, b).

If (a, b) is a point in the domain of F, and if $F(x, y) \geqslant F(a, b)$ for all (x, y) in the domain of F sufficiently close to (a, b), then we say that F has a local minimum at (a, b).

The difficulty with the above definitions is that they depend on the meaning of "sufficiently close", and it is this phrase which needs to be precisely defined.

If we use our approach to functions of one variable as a guide, then we need a "small" set in $R \times R$ where before we had a "small" interval, $[c - \varepsilon, c + \varepsilon]$ in R; the most suitable set in $R \times R$ is a circular disc.

We let $S(a, b, \varepsilon)$ denote the set $\{(x, y):(x - a)^2 + (y - b)^2 \leqslant \varepsilon^2\}$, which is a disc with centre at the point (a, b) and radius ε.
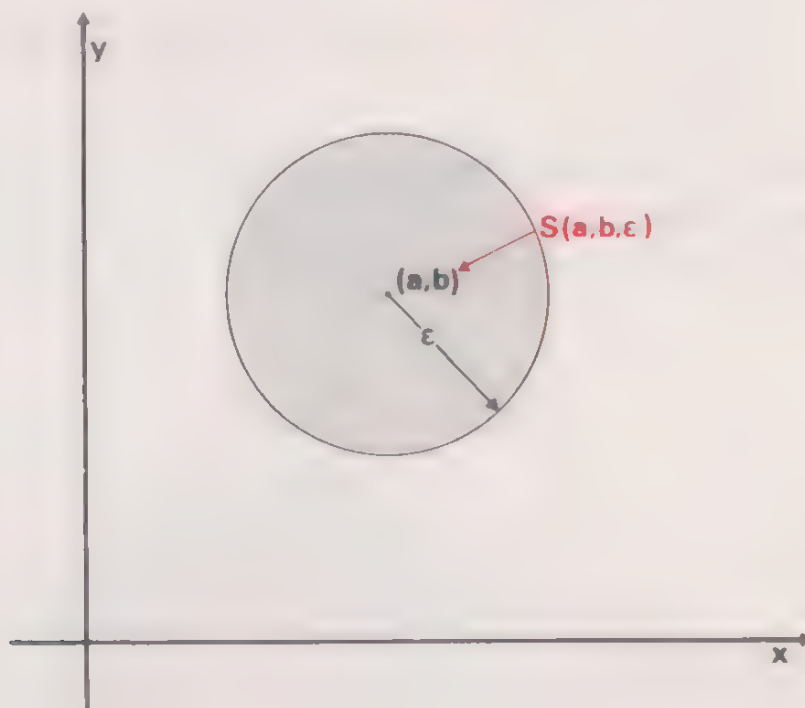
Let F be a function with domain $A \subseteq R \times R$. Then, following our definitions for functions of one variable, we make the following formal definitions.

If there is a positive number ε such that $F(x, y) \leqslant F(a, b)$ for all $(x, y) \in A \cap S(a, b, \varepsilon)$, then we say that F has a local maximum at (a, b).

If there is a positive number ε such that $F(x, y) \geqslant F(a, b)$ for all $(x, y) \in A \cap S(a, b, \varepsilon)$, then we say that F has a local minimum at (a, b).

Let us have a look at the set $A \cap S(a, b, \varepsilon)$. Take the set in the following diagram to be the set A.
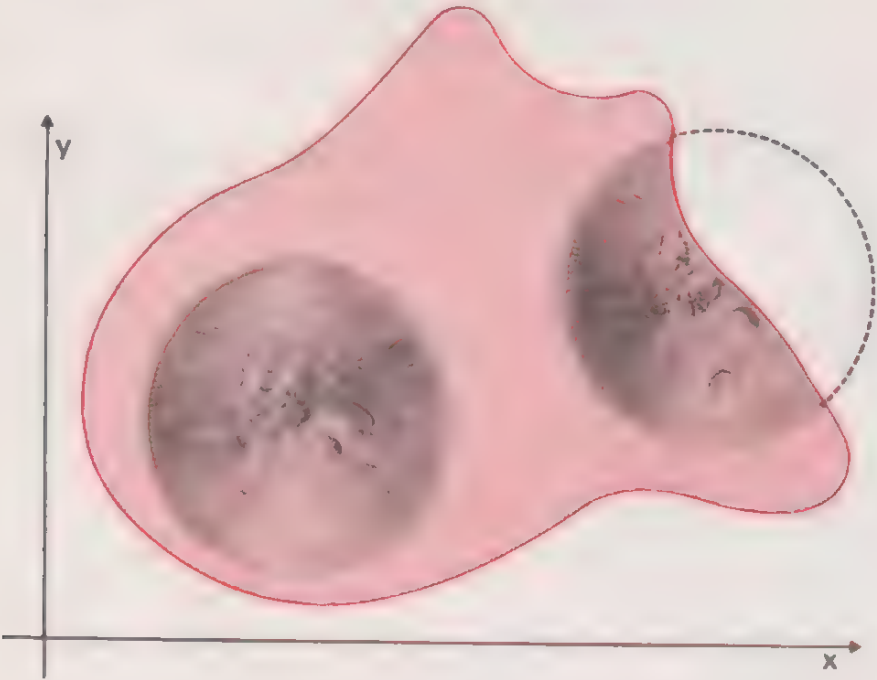
If we represent S by a 10p coin, then, placing the 10p down on the set, the part of A which is covered by the coin is the set $A \cap S(a, b, \varepsilon)$. The point (a, b) is of course the centre point of the coin, and ε is its radius.

Example 1

(i)

A

$A \cap S(a, b, \varepsilon)$
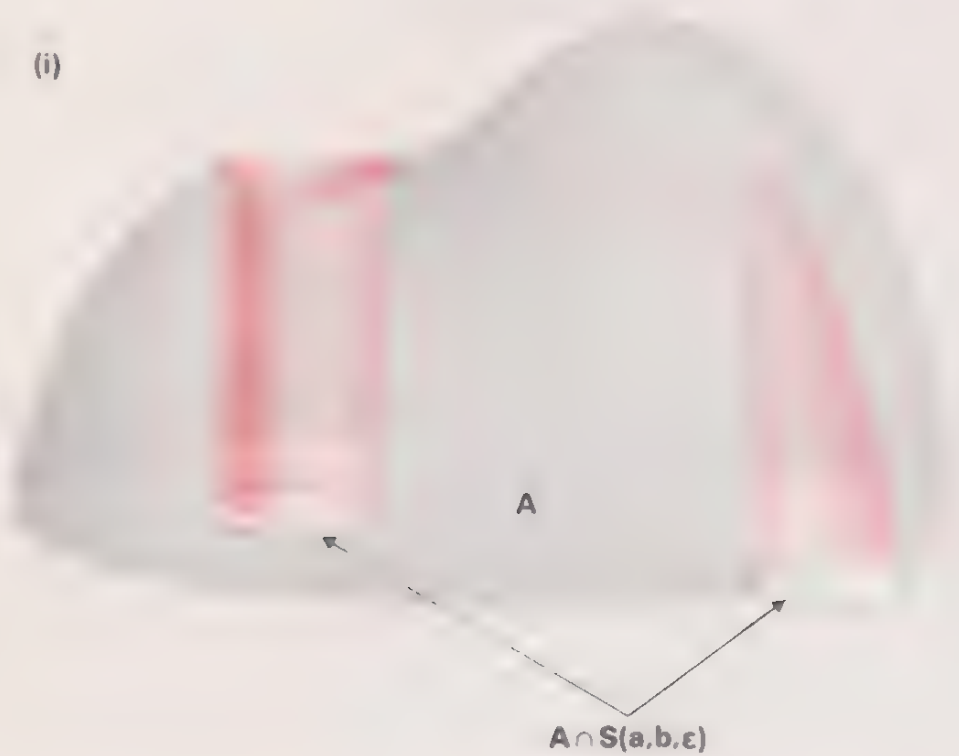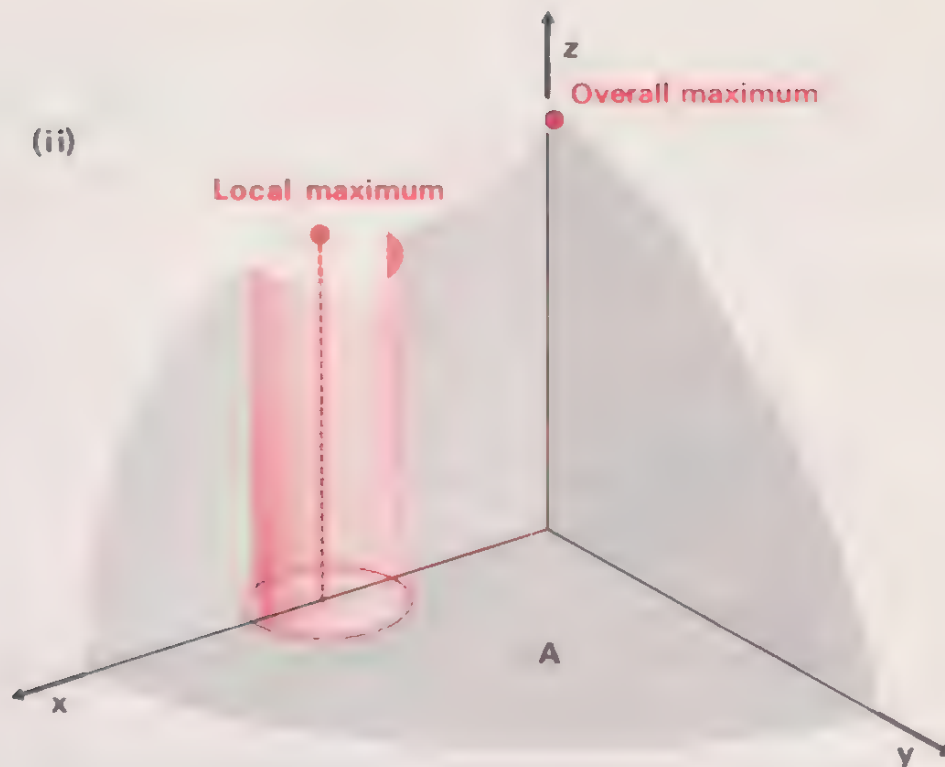
Just as for functions of one variable, we have two major problems. A stationary point need not be a local maximum nor a local minimum. Stationary points of this kind are called saddle points (for obvious reasons).

Horizontal tangent plane at a saddle point

In other words, saying that the tangent plane is horizontal guarantees neither a local maximum nor a local minimum.

The second big problem is that a local maximum, or indeed an overall maximum, can occur on the boundary of the domain, and similarly for local and overall minima. If we restrict ourselves to a search for stationary points, then we may miss points of this kind.

We shall attempt to overcome the first problem, but we shall not overcome the second problem in this volume. In general, the second problem is not so serious when applying the methods to physical situations, since the information from the situation itself often helps us to decide the nature of the point we are dealing with.

A Useful Technique

Example 2

Find the overall minimum value of the function

$$F:(x, y) \longmapsto x^2 + 2y^2 \qquad ((x, y) \in R \times R).$$

Since both expressions on the right are positive or zero, the answer is obviously 0, and this occurs when $x = y = 0$.

Now let us use the above example to test a technique for classifying stationary points, which we can use when the answer isn't obvious.

The surface $z = x^2 + 2y^2$ looks like this:
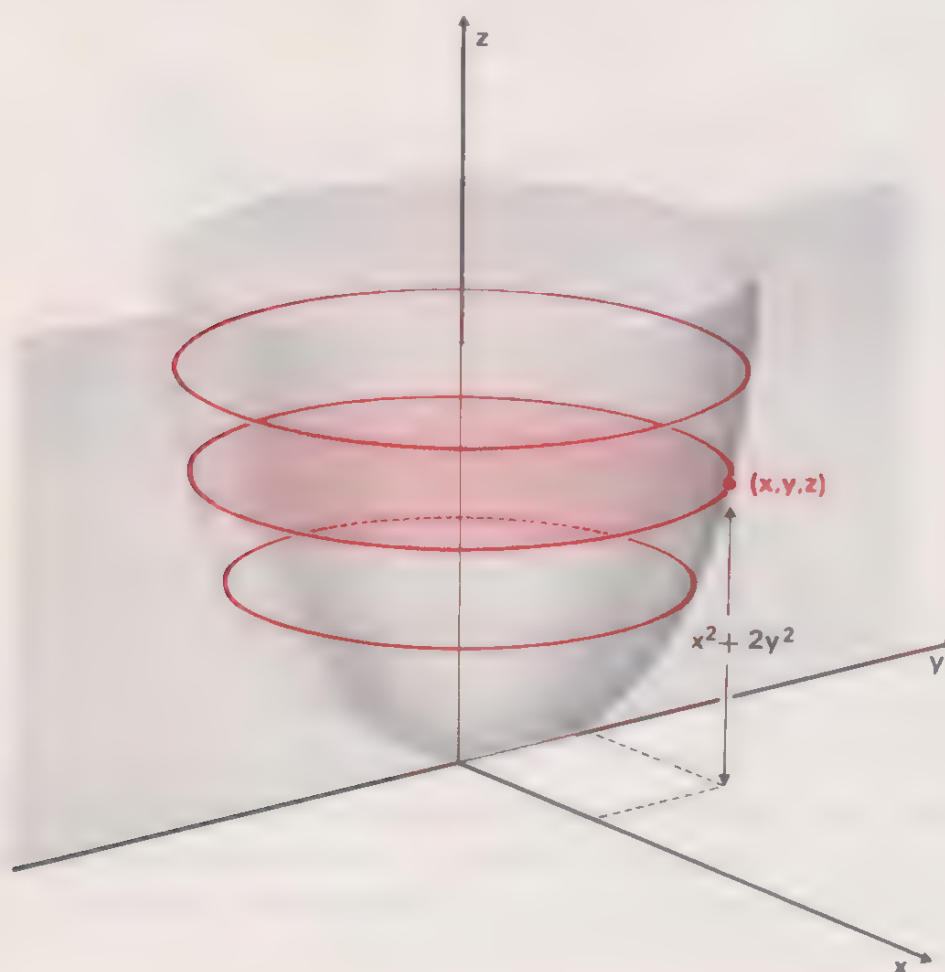
Notice that

$$F'_1(a, b) = 2a \quad \text{and} \quad F'_2(a, b) = 4b,$$

and therefore the point corresponding to $a = b = 0$ is a stationary point, confirming what we already know.

The essential idea of our proposed technique is as follows: intersect the surface with the plane whose equation is

$$y = x \tan \alpha,$$

to give the curve shown in red.
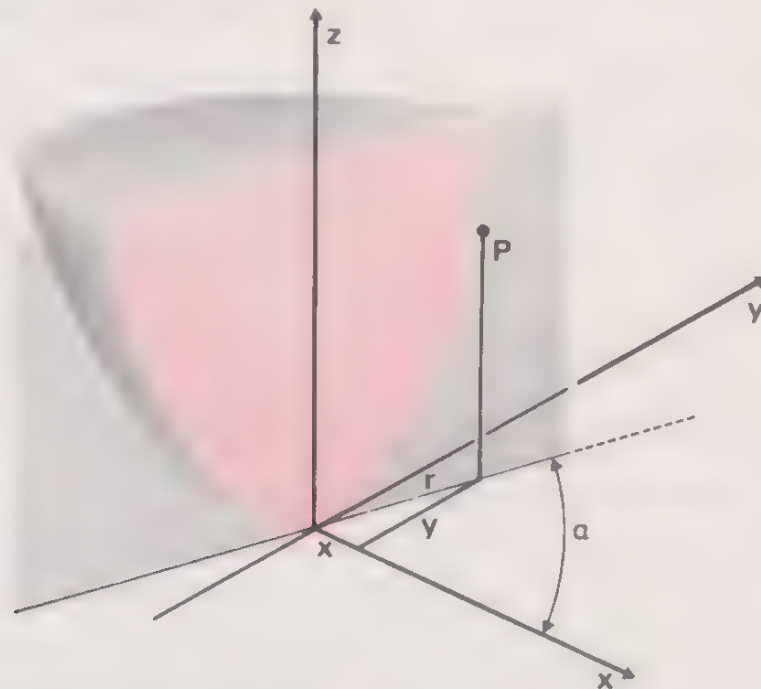
It is obvious from the diagram that the red curve has a minimum at 0, but can we show that this is the case mathematically (thus making the diagram redundant)?

The red curve is determined geometrically as the intersection of the plane, $\{(x, y) : y = x \tan \alpha\}$, and the surface, $\{(x, y) : z = x^2 + 2y^2\}$, or, more briefly, it is determined by the equations

$$z = x^2 + 2y^2$$

$$y = x \tan \alpha.$$

If r is the hypotenuse of the right-angled triangle with sides x and y (shown on the previous diagram), then the second equation can be replaced by

$$x = r \cos \alpha \quad \text{and} \quad y = r \sin \alpha.$$

On the red curve we then have

$$z = r^2(\cos^2 \alpha + 2 \sin^2 \alpha)$$
$$= r^2(1 + \sin^2 \alpha),$$

so that on the red curve, $z(= F(x, y))$ takes its minimum value when $r = 0$.

In more difficult cases we might have to adopt the following line of reasoning to achieve the required result.

The equation $z = r^2(1 + \sin^2 \alpha)$ defines a function of one variable:

$$\phi : r \longmapsto z \quad (r \in R_0^+, \text{ the set of non-negative real numbers}).$$

Differentiating, we obtain

$$\phi'(r) = 2r(1 + \sin^2 \alpha)$$

and

$$\phi''(r) = 2(1 + \sin^2 \alpha).$$

In particular, $\phi''(0) = 2(1 + \sin^2 \alpha)$, so that $\phi''(0) > 0$, and therefore the red curve has a *local minimum* at $r = 0$.

The essential point about the equation $\phi''(0) = 2(1 + \sin^2 \alpha)$ is that it shows that $\phi''(0) > 0$ *for all values of* α, so that all possible red curves obtained in this way have a local minimum at 0. This seems to show pretty conclusively that F has a local minimum at 0 (again confirming the result which we already know).

Roughly speaking, the technique can be summarized as follows. Slice through the point of interest on the surface with a vertical cut, revealing a curve like the red curve of our example. If all such curves have a local minimum at the point, then we would expect the surface to have a local minimum there. Our next example will again illustrate this idea.

This technique for classifying the stationary points will be adequate when the surface is "smooth", and it will certainly be adequate for all the problems which you will meet in this volume. However, it is an amazing fact that one can construct a surface for which all the red curves have a local minimum at the origin, *and yet* the surface does *not* have a local minimum at that point. Such a surface cannot be "smooth" in our sense, and you may like to try to think of such an example.

Example 3 (This example has applications in statistics.)

Given the three points with co-ordinates (1, 1), (2, 3) and (3, 4), find a line with equation $y = mx + c$, such that the sum of the squares of the "vertical" distances of the points from the line is a minimum.
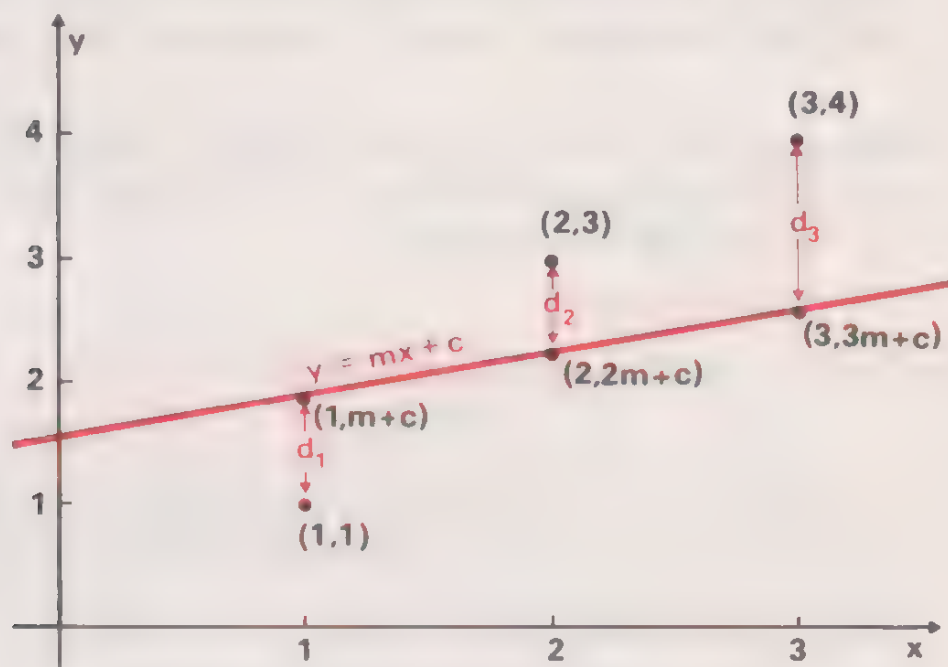
In the diagram, d_1, d_2 and d_3 are the vertical distances, and we want to minimize

$$d_1^2 + d_2^2 + d_3^2 = (m + c - 1)^2 + (2m + c - 3)^2 + (3m + c - 4)^2;$$

we can therefore define a function F with domain $R \times R$ by putting

$$F(m, c) = (m + c - 1)^2 + (2m + c - 3)^2 + (3m + c - 4)^2.$$

<div align="right">**Equation (1)**</div>

We then have

$$F_1'(m, c) = 2(m + c - 1) + 4(2m + c - 3) + 6(3m + c - 4)$$

$$= 28m + 12c - 38$$

and

$$F_2'(m, c) = 2(m + c - 1) + 2(2m + c - 3) + 2(3m + c - 4)$$

$$= 12m + 6c - 16.$$

The values of m and c for which $F_1'(m, c) = F_2'(m, c) = 0$ are determined by the equations

$$14m + 6c - 19 = 0$$

$$12m + 6c - 16 = 0$$

from which we deduce that

$$m = \tfrac{3}{2} \quad \text{and} \quad c = -\tfrac{1}{3}.$$

These values of m and c determine a stationary point of F, but is this stationary point the overall minimum? Could we show that the stationary point is even a local minimum? Let us try the method of our previous example.

Our variables are m and c instead of x and y, and so we intersect the unknown surface by planes perpendicular to the mc-plane and through the point $(\tfrac{3}{2}, -\tfrac{1}{3})$. A typical plane has equation

$$c + \tfrac{1}{3} = (m - \tfrac{3}{2}) \tan \alpha$$

and just as before we can replace this equation by

$$m = \tfrac{3}{2} + r \cos \alpha \qquad c = -\tfrac{1}{3} + r \sin \alpha$$

where r is now the distance of the point with co-ordinates (m, c) from the point $(\tfrac{3}{2}, -\tfrac{1}{3})$. Substituting these expressions for m and c into Equation (1), we obtain the function ϕ defined by

$$\phi(r) = (r(\cos \alpha + \sin \alpha) + \tfrac{1}{6})^2 + (r(2 \cos \alpha + \sin \alpha) - \tfrac{1}{3})^2$$

$$+ (r(3 \cos \alpha + \sin \alpha) + \tfrac{1}{6})^2 \qquad (r \in R_0^+).$$

Differentiating twice

$$\phi''(0) = 2((\cos \alpha + \sin \alpha)^2 + (2 \cos \alpha + \sin \alpha)^2$$

$$+ (3 \cos \alpha + \sin \alpha)^2)$$

and therefore $\phi''(0) > 0$ for *all* values of α (since the squares cannot be zero simultaneously); hence the stationary point is a local minimum.

Since there is only one stationary point and it is a local minimum, it seems very likely that we have indeed found the required values of m and c. So the equation of the required line is $6y = 9x - 2$.

The difficulty with points on the boundary of the domain of F does not occur in this case, because the domain is the whole set $R \times R$, and there are no boundary points; but to be safe we ought really to find the images of the function when r is very large. For the moment we shall avoid this difficulty too.

Exercise 1

Find the equation of the red line which gives the minimum value of $p_1^2 + p_2^2 + p_3^2$.
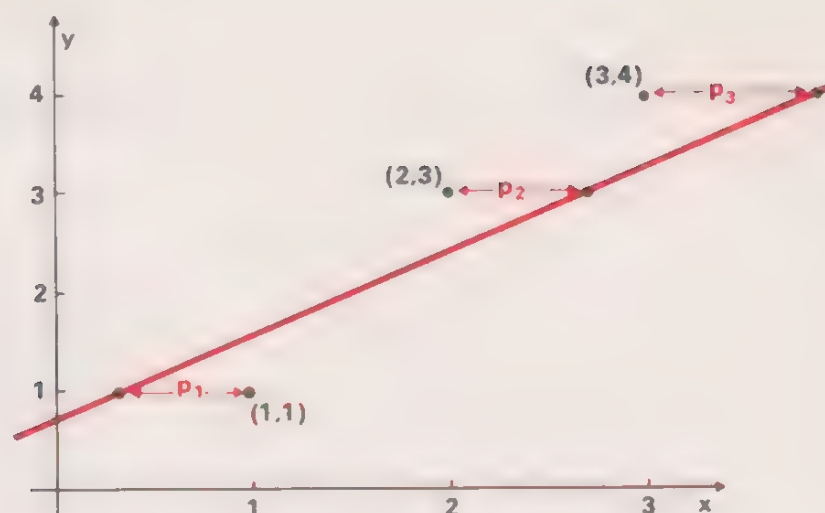
Writing the equation of the line in the form $y = mx + c$ leads to some untidy algebra. In this exercise it is convenient to take $x = my + c$ as the equation of the line.

2.6 Additional Exercises

Exercise 1

Find the partial derivatives at (x, y) of the functions defined by the following equations; each function has domain $R \times R$.

(i) $F(x, y) = x \sin (x + y)$
(ii) $G(x, y) = x^4 + y^4 - 4x^2y^3$

Exercise 2

Find the equation of the tangent plane at the point on the surface corresponding to the pair (a, b) for each of the following functions (having domain $R \times R$):

(i) $F: (x, y) \longmapsto x \sin (x + y)$
(ii) $G: (x, y) \longmapsto x^4 + y^4 - 4x^2y^3$

(Use the results of Exercise 1 above.)

Exercise 3

Find the equation of the tangent plane to the surface defined by

$$F : (x, y) \longmapsto (4 - 2x + y)\sqrt{x^2 - y^2}$$

$$((x, y) \in \{(x, y) : 0 \leqslant y \leqslant x \leqslant 2\})$$

at the point $(a, b, F(a, b))$.

Exercise 4

Find the point P for which the sum $S_1^2 + S_2^2 + S_3^2$ is a minimum.
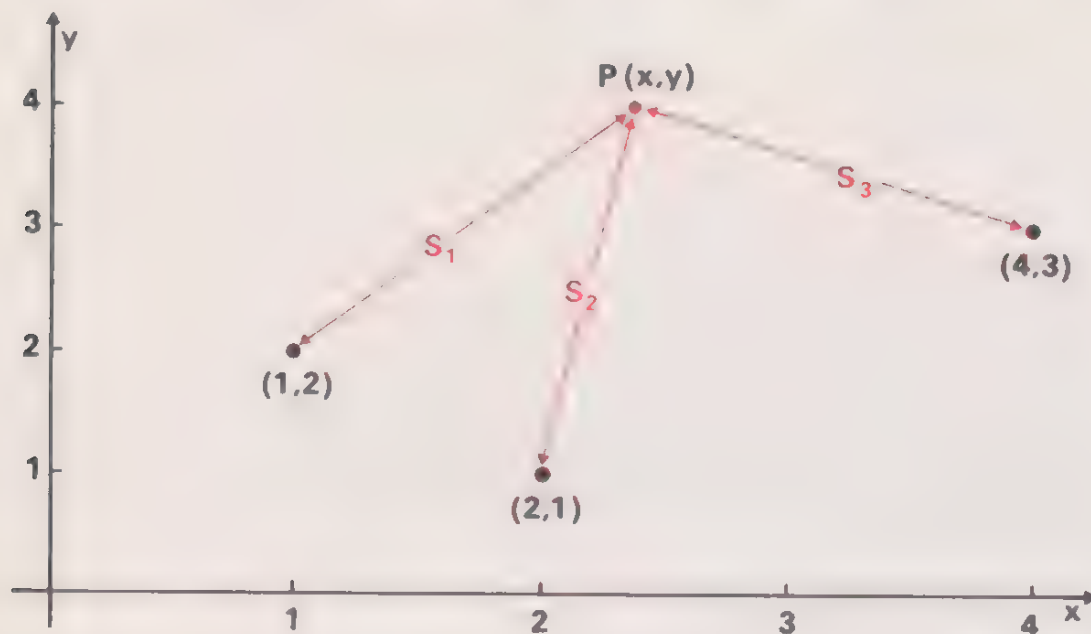
2.7   Answers to Exercises

Section 2.1

Exercise 1

(i)-(iii)
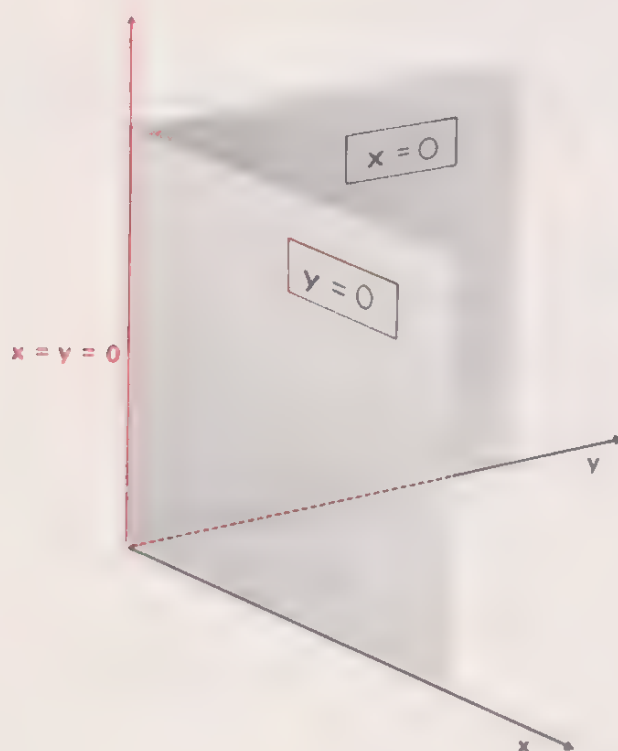
Exercise 2

Considering the equation as a restriction defining a subset of $R \times R \times R$, we see that z does not appear in the equation, so there is no restriction on z. But x and y are restricted. If (x, y, z) is to belong to the subset, then x and y must satisfy the equation $2x - y = 0$. The set of triples $(x, y, 0)$ which satisfy this equation form a line in the xy-plane.
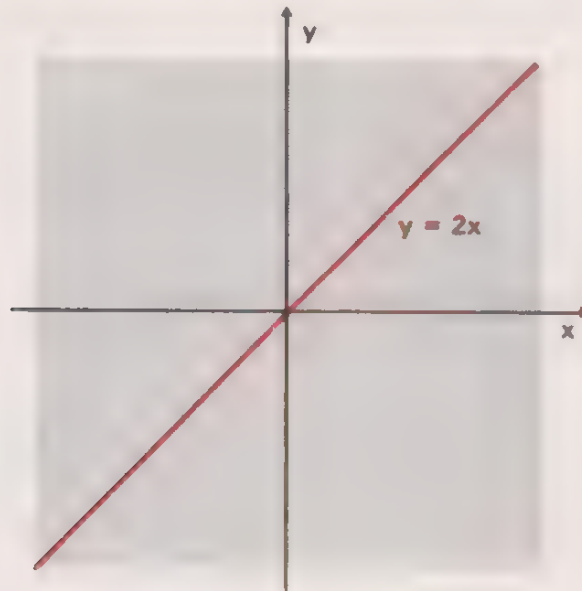
Corresponding to any point on this line, we can get other elements of the required subset of $R \times R \times R$ by choosing any value of z. All in all, we get a plane perpendicular to the xy-plane, which intersects the xy-plane in the line with equation $2x - y = 0$.
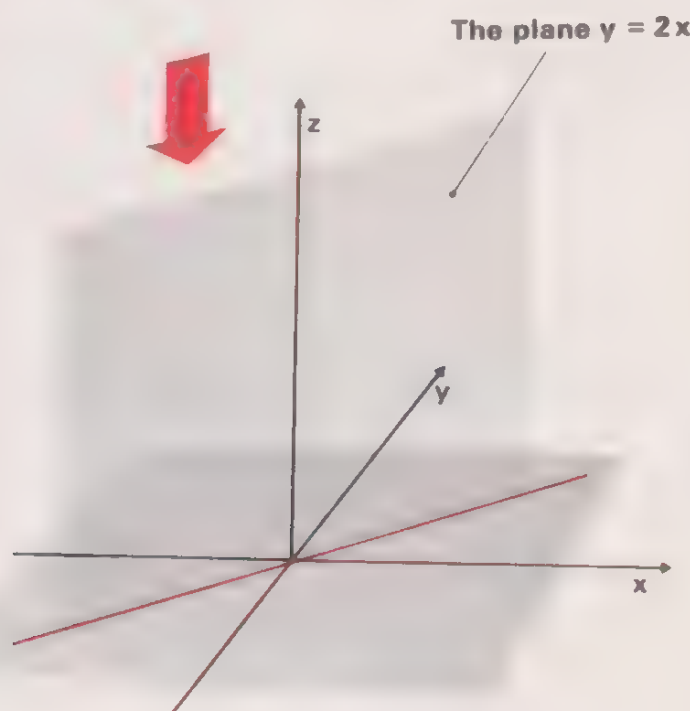
Section 2.2

Exercise 1

(i) $\alpha = \tan A$
$\beta = \tan B$
$\gamma = -1$
$\delta = c - a \tan A - b \tan B.$

(ii) If $\alpha \neq 0$, the plane meets the x-axis in the point with co-ordinates
$\left(-\dfrac{\delta}{\alpha}, 0, 0\right)$.

If $\alpha = 0$, the equation of the plane is $\beta y + \gamma z + \delta = 0$:

(a) If $\delta \neq 0$, then y and z cannot be zero simultaneously, so the plane does not meet the x-axis.

(b) If $\delta = 0$, the equation of the plane is $\beta y + \gamma z = 0$. All points with co-ordinates $(x, 0, 0)$ lie on this plane, and so the plane contains the whole x-axis.

A similar argument can be used for the other co-ordinate axes.

(iii) Varying λ and μ alters the line in which the plane cuts the xy-plane. It will not change the fact that the plane is perpendicular to the xy-plane.

Section 2.3

Exercise 1

(i) $F_1'(x, y) = 2x$
$F_2'(x, y) = 2y$

(ii) $G_1'(x, y) = \exp(xy) + xy \exp(xy)$
$G_2'(x, y) = x^2 \exp(xy)$

Section 2.4

Exercise 1

(i) $z = a^2 + b^2 + 2a(x - a) + 2b(y - b)$

(ii) $z = a \exp(ab) + (ab \exp(ab) + \exp(ab))(x - a) + a^2 \exp(ab)(y - b)$

Section 2.5

Exercise 1

$$p_1^2 + p_2^2 + p_3^2 = (1 - (m + c))^2 + (2 - (3m + c))^2 + (3 - (4m + c))^2$$

$$= 26m^2 + 16mc + 3c^2 - 38m - 12c + 14.$$

So F is the function defined by

$$F:(m, c) \longmapsto 26m^2 + 16mc + 3c^2 - 38m - 12c + 14$$

$$((m, c) \in R \times R)).$$

Thus, for a stationary point, we have

$$F'_1(m, c) = 52m + 16c - 38 = 0$$

$$F'_2(m, c) = 16m + 6c - 12 = 0.$$

The solution of this pair of equations is

$$m = \tfrac{9}{14}, \qquad c = \tfrac{2}{7}.$$

An argument similar to that given in the text would show that these values of m and c do give a local minimum. Thus the equation of the line is

$$14x = 9y + 4.$$

Section 2.6

Exercise 1

(i) $F'_1(x, y) = \sin(x + y) + x \cos(x + y)$

$\quad F'_2(x, y) = x \cos(x + y)$

(ii) $G'_1(x, y) = 4x^3 - 8xy^3$

$\quad G'_2(x, y) = 4y^3 - 12x^2y^2$

Exercise 2

(i) $\quad z = a \sin(a + b) + (\sin(a + b) + a \cos(a + b))(x - a)$
$\qquad + a \cos(a + b)(y - b)$

(ii) $\quad z = a^4 + b^4 - 4a^2b^3 + (4a^3 - 8ab^3)(x - a)$
$\qquad + (4b^3 - 12a^2b^2)(y - b).$

Exercise 3

First, we must find the partial derivatives; they are:

$$F'_1(x, y) = -2\sqrt{x^2 - y^2} + (4 - 2x + y)\frac{x}{\sqrt{x^2 - y^2}}$$

$$F'_2(x, y) = \sqrt{x^2 - y^2} - (4 - 2x + y)\frac{y}{\sqrt{x^2 - y^2}}$$

Thus the equation of the tangent plane is

$$z = (4 - 2a + b)\sqrt{a^2 - b^2}$$

$$+ \left(-2\sqrt{a^2 - b^2} + \frac{(4 - 2a + b)}{\sqrt{a^2 - b^2}}a \right)(x - a)$$

$$+ \left(\sqrt{a^2 - b^2} - \frac{(4 - 2a + b)}{\sqrt{a^2 - b^2}}b \right)(y - b)$$

Exercise 4

The sum

$$S_1^2 + S_2^2 + S_3^2 = (x - 1)^2 + (y - 2)^2 + (x - 2)^2$$

$$+ (y - 1)^2 + (x - 4)^2 + (y - 3)^2$$

$$= 3x^2 + 3y^2 - 14x - 12y + 35.$$

So we let F be the function

$$F : (x, y) \longmapsto 3x^2 + 3y^2 - 14x - 12y + 35 \qquad ((x, y) \in R \times R).$$

We have

$$F_1'(x, y) = 6x - 14$$

$$F_2'(x, y) = 6y - 12.$$

Thus for a stationary point, $x = 2\frac{1}{3}$ and $y = 2$.

Once again we can use either the argument given in the text or the geometry of the situation to convince ourselves that $(2\frac{1}{3}, 2)$ does in fact give the point P for which the sum is a minimum.

CHAPTER 3   TECHNIQUES OF INTEGRATION

3.0   Introduction

In Volume 1, Chapters 7 and 9, we introduced the basic ideas of *integration* and we used *the fundamental theorem of calculus* to show how integration and differentiation are related.

In this chapter, we return to the topic of *integration* and we introduce two powerful techniques for evaluating integrals. These are based on the following formula which we obtained in Chapter 9 of Volume 1:

$$\int_a^b DF = F(b) - F(a),$$

or on the equivalent statement about primitive functions:

F is a primitive function of DF.

3.1   Integration by Parts

In this section we formulate the rule of integration that corresponds to the rule for differentiating the product of two functions. It is useful when dealing with integrals of products of functions.

The rule for differentiating a product of two real functions f and g obtained in Volume 1, Chapter 8, is

$$D(f \times g) = f \times Dg + g \times Df$$

where $\times$ denotes the multiplication of functions. To convert this into a rule for integration, we take a definite integral of both sides, obtaining

$$\int_a^b D(f \times g) = \int_a^b (f \times Dg) + \int_a^b (g \times Df)$$

where a and b are any numbers such that $[a, b]$ is included in the domains of the functions Df and Dg. Applying the Fundamental Theorem of Calculus, we can put this equation into the form:

$$[f \times g]_a^b = \int_a^b (f \times Dg) + \int_a^b (g \times Df)$$

or, on rearranging,

$$\int_a^b (f \times Dg) = [f \times g]_a^b - \int_a^b (g \times Df).$$

This is called the rule for integration by parts, because we integrate only *part* of the function under the integral sign on the left — the part Dg.

At first sight the rule of integration by parts does not look as though it will help much in the evaluation of integrals, because it converts one integral, $\int_a^b (f \times Dg)$, into an apparently more complicated expression that involves another integral looking very much like the one with which we started. When specific functions are used in place of the unspecified functions f and g, however, it may happen that the new integral, $\int_a^b (g \times Df)$, is easier to evaluate than the old one, $\int_a^b (f \times Dg)$, and if so, then the rule of integration by parts will have served its purpose.

As an illustration, we apply the method to the integral

$$\int_a^b x \longmapsto x \cos x.$$

The function to be integrated is the product of the function $x \longmapsto x$ and the function $x \longmapsto \cos x$, which can be abbreviated to cos, so the integral is

$$\int_a^b (x \longmapsto x) \times \cos.$$

The rule of integration by parts is

$$\int_a^b (f \times Dg) = [f \times g]_a^b - \int_a^b (g \times Df)$$

and to use it on our integral, we take f to be $x \longmapsto x$ and Dg to be cos. From the table of standard integrals we know that $D \sin = \cos$, so we take g to be the function sin. (Any other primitive function of cos would do instead, but the one used here is the natural choice, because it is the simplest.) With these choices for f and g we have

$$f : x \longmapsto x, \qquad Df : x \longmapsto 1$$

and

$$g : x \longmapsto \sin x, \qquad Dg : x \longmapsto \cos x$$

and so our integral becomes

$$\int_a^b (x \longmapsto x) \times \cos = [(x \longmapsto x) \times \sin]_a^b - \int_a^b \sin \times (x \longmapsto 1)$$

which means the same as

$$\int_a^b x \longmapsto x \cos x = [x \longmapsto x \sin x]_a^b - \int_a^b x \longmapsto \sin x.$$

The integral on the right is easier to evaluate than the one on the left; in fact it is a standard integral, and $x \longmapsto -\cos x$ is a primitive function. We can therefore evaluate the right-hand side, obtaining the required integral:

$$\int_\blacksquare^b x \longmapsto x \cos x = [x \longmapsto x \sin x]_a^b - [x \longmapsto -\cos x]_a^b$$

$$= b \sin b - a \sin a + \cos b - \cos a.$$

When you come to apply the formula for integration by parts you may find it more convenient to use the following form, in which the images of the functions are shown explicitly:

$$\int_a^b x \longmapsto f(x) \times \mathbf{D}g(x)$$

$$= [x \longmapsto f(x) \times g(x)]_a^b - \int_a^b x \longmapsto g(x) \times Df(x).$$

Remember that $Dg(x)$ is the same thing as $g'(x)$, the image of x under the derived function Dg, which is the derivative of g at x. For example, this formula, when applied to the integral we have just treated, gives

$$\int_a^b x \longmapsto x \cos x = [x \longmapsto x \sin x]_a^b - \int_a^b \sin x \times (x \longmapsto 1)$$

where

$$f(x) = x, \qquad Df(x) = 1,$$

$$g(x) = \sin x, \qquad Dg(x) = \cos x.$$

The notation is, however, still clumsy, and we suggest that where:

  (i) the functions have been clearly defined at the beginning of a piece of work;

 (ii) there is no likelihood of confusion; for example, it is quite clear which symbol is being used for the variable defining the function;

(iii) the complexity of the work warrants it;

the "$x \longmapsto$" part of the notation of a function be dropped for the purposes of calculation. This is, of course, an "abuse of notation"; but

it is normal mathematical practice to abuse notation when the situation warrants it.

Modifying our notation, we obtain:

$$\int_a^b x \cos x = [x \sin x]_a^b - \int_a^b \sin x \times 1,$$

and the general formula for integration by parts becomes

$$\int_a^b f(x) \times Dg(x) = [f(x) \times g(x)]_a^b - \int_a^b g(x) \times Df(x).$$

Another useful piece of notation is the following. So far we have denoted one of the primitive functions of a given function f by the corresponding capital letter F. This now becomes inconvenient, because $F \times DG$ is not a primitive of $f \times Dg$, so we denote one of the primitive functions of a given function f by

$$\int f$$

that is, we use the integration symbol without the end-points of integration. In terms of primitive functions, the formula for integration by parts becomes

$$\int f \times Dg = f \times g - \int g \times Df,$$

where the two primitive functions in this formula will be determined by their context: the result asserts that, if $\int g \times Df$ is one of the primitive functions of $g \times Df$, then $f \times g - \int g \times Df$ is one of the primitive functions of $f \times Dg$. For example, if we are asked to find a primitive function of $x \longmapsto x \exp x \ (x \in R)$, then we choose

$$f : x \longmapsto x, \qquad g : x \longmapsto \exp x,$$

and obtain

$$\int x \exp x = x \exp x - \int \exp x \times 1$$

$$= x \exp x - \exp x$$

$$= (x - 1) \exp x$$

that is, one of the primitive functions of $x \longmapsto x \exp x$ is

$$x \longmapsto (x - 1) \exp x.$$

Exercise 1

Evaluate $\displaystyle\int_0^\pi x \longmapsto x \sin x.$

(HINT: Take $f(x) = x$ in the formula for integration by parts.)

Exercise 2

Apply the rule of integration by parts twice in succession to find a primitive function of $x \longmapsto x^2 \exp x \ (x \in R)$.

(HINT: take $f(x) = x^2$ in the first integration by parts.)

3.2   Integration by Substitution

One can often evaluate an integral of a function most easily by finding a primitive which is a composition of more elementary functions. First of all we must find out how to integrate a composite function. As an illustration, consider the problem of evaluating

$$\int_a^b x \longmapsto x \cos (x^2)$$

where a and b are positive real numbers. This looks very similar to the integral which we evaluated by parts in the preceding section, but the fact that the integrand now involves $\cos (x^2)$ instead of $\cos x$ makes a big difference. If we try to apply the method that we used for

$$\int_a^b x \longmapsto x \cos x,$$

we find that the functions f and g enter the calculation, where

$$f(x) = x \qquad Dg(x) = \cos (x^2)$$

$$Df(x) = 1, \qquad g(x) = ?$$

Before, we had $Dg(x) = \cos x$, so that $g(x)$ was $\sin x$; but in this case there is no simple function having derived function $x \longmapsto \cos (x^2)$ to use for g. This is not the only way of using the rule of integration by parts here, but none of the alternatives is much help either; so instead of labouring the

integration by parts method any more, let us look instead at the integral

$$\int_a^b x \longmapsto x \cos(x^2)$$

from a fresh point of view.

One way to evaluate the integral would be to find a suitable primitive function, and by the Fundamental Theorem of Calculus this primitive function, F say, will satisfy the equation:

$$DF(x) = x \cos(x^2) \qquad (x \in [a, b]).$$

<div align="right">**Equation (1)**</div>

The expression $\cos(x^2)$ suggests that F may have the form

$$F(x) = G(x^2) \qquad (x \in [a, b])$$

<div align="right">**Equation (2)**</div>

where G is some new function, to be chosen in accordance with Equation (1). To use this equation we differentiate the function in Equation (2), obtaining

$$DF(x) = 2x \, DG(x^2) \qquad (x \in [a, b])$$

by the rule for differentiating composite functions (see Volume I, Chapter 8); Equation (1) then gives:

$$2x \, DG(x^2) = x \cos(x^2) \qquad (x \in [a, b]).$$

The natural way to satisfy this condition is to make

$$DG(x^2) = \tfrac{1}{2} \cos(x^2) \qquad (x \in [a, b]),$$

that is,

$$DG(u) = \tfrac{1}{2} \cos u \qquad (u \in [a^2, b^2])$$

<div align="right">**Equation (3)**</div>

where u stands for x^2. We have now reduced the problem of finding a primitive of $x \longmapsto x \cos(x^2)$ to the simpler one of finding a primitive function of $u \longmapsto \tfrac{1}{2} \cos u$.

By the table of standard integrals, this latter primitive function is $\tfrac{1}{2} \sin$, so from Equation (3) we obtain:

$$G(u) = \tfrac{1}{2} \sin u \qquad (u \in [a^2, b^2]),$$

and then Equation (2) gives a required primitive function F, where

$$F(x) = \tfrac{1}{2} \sin(x^2) \qquad (x \in [a, b]).$$

The integral we set out to evaluate is therefore

$$\int_a^b x \cos (x^2) = [\tfrac{1}{2} \sin (x^2)]_a^b$$

$$= \tfrac{1}{2} \sin (b^2) - \tfrac{1}{2} \sin (a^2).$$

Exercise 1

Write down the appropriate entries for the empty boxes and hence evaluate:

$$\int_{\pi^2/4}^{\pi^2} x \longmapsto \frac{\sin \sqrt{x}}{\sqrt{x}}.$$

If

$$F(x) = G(\sqrt{x}) \qquad \left(x \in \left[\frac{\pi^2}{4}, \pi^2\right]\right),$$

then we have

$$DF(x) = \boxed{} DG(\sqrt{x}). \qquad \text{(i)}$$

Putting

$$DF(x) = \frac{\sin \sqrt{x}}{\sqrt{x}}, \quad \text{we get}$$

$$\sin \sqrt{x} = \boxed{} . \qquad \text{(ii)}$$

If we put

$$\sqrt{x} = u, \quad \text{then}$$

$$DG(u) = \boxed{} \qquad \text{(iii)}$$

and a primitive function for $u \longmapsto 2 \sin u$ is

$$u \longmapsto \boxed{} . \qquad \text{(iv)}$$

Thus we find

$$G(u) = \boxed{}.$$

(v)

The definition of G gives:

$$G(u) = G(\sqrt{x}) = F(x).$$

Thus we can find $F(x)$ in terms of x. The expression is

$$F(x) = \boxed{}$$

(vi)

and since

$$\int_{\pi^2/4}^{\pi^2} x \longmapsto \frac{\sin\sqrt{x}}{\sqrt{x}} = [F]_{\pi^2/4}^{\pi^2}$$

we have

$$\int_{\pi^2/4}^{\pi^2} x \longmapsto \frac{\sin\sqrt{x}}{\sqrt{x}} = \boxed{}$$

(vii)

Exercise 2

By writing $F(x) = G(-x^2)$, and making a suitable choice for F, evaluate

$$\int_a^b x \longmapsto x\exp(-x^2),$$

where a and b are any positive real numbers such that $a < b$.

To make the application of this method as convenient as possible, it is worth setting up a general formula, just as we did for integration by parts, embodying the steps that are common to every application of the method. The method we have been discussing in this section comes from the result for differentiating a composite function. For example, in the evaluation of

$$\int_a^b x \longmapsto x\cos(x^2)$$

we looked for a primitive function F in the form

$$F(x) = G(x^2)$$

(see Equation (2)). In general, this composite function has the form

$$F = G \circ k,$$

in the notation introduced in Volume 1, Chapter 3, so

$$F(x) = G(k(x)).$$

In general, if the integral we are trying to evaluate is $\int_a^b f$, then the Fundamental Theorem of Calculus tells that $f = DF$, and hence, by the rule for differentiating composite functions we have:

$$F' = (G' \circ k) \times k'$$

that is,

$$f = DF = (DG \circ k) \times Dk.$$

Thus the integral we are trying to evaluate has the form

$$\int_a^b f = \int_a^b (DG \circ k) \times Dk \qquad\qquad \text{Equation (4)}$$

and its value is

$$F(b) - F(a) = G(k(b)) - G(k(a)). \qquad\qquad \text{Equation (5)}$$

Since Equation (4) will not give us G directly, but will give us DG (if we know k), it is best to express Equation (5) in terms of DG too; by the Fundamental Theorem we can put Equation (5) in the form:

$$F(b) - F(a) = \int_{k(a)}^{k(b)} DG. \qquad\qquad \text{Equation (6)}$$

Writing g for DG and combining Equations (4) and (6), we get:

$$\int_a^b (g \circ k) \times Dk = \int_{k(a)}^{k(b)} g.$$

This is the basic rule for integration by substitution.

Example 1

As an example, we apply the rule to the integral we considered at the beginning of this section, that is:

$$\int_a^b x \longmapsto x \cos(x^2).$$

Our previous calculation corresponds to the choice

$$k(x) = x^2 \qquad\qquad (x \in R).$$

Since $Dk(x) = 2x$, and we require

$$(g \circ k) \times Dk = x \longmapsto x \cos(x^2),$$

we take

$$g \circ k(x) = \tfrac{1}{2} \cos(x^2) \qquad (x \in R)$$

which is the equivalent to

$$g(u) = \tfrac{1}{2} \cos u \qquad (u \in R, \text{ and } u \geqslant 0)$$

where we have written u for $k(x)$, that is, for x^2.

This substitution of u for $k(x)$ greatly simplifies the manipulations, and accounts for the name "integration by substitution". The rule now tells us that:

$$\int_a^b x \longmapsto x \cos(x^2) = \int_{k(a)}^{k(b)} g$$

$$= \int_{a^2}^{b^2} u \longmapsto \tfrac{1}{2} \cos u$$

$$= [\tfrac{1}{2} \sin u]_{a^2}^{b^2}$$

$$= \tfrac{1}{2} \sin(b^2) - \tfrac{1}{2} \sin(a^2)$$

as we found before.

Sometimes the rule of substitution is most conveniently used in a "backwards" form in which we start from the right-hand side of the basic formula

$$\int_a^b (g \circ k) \times Dk = \int_{k(a)}^{k(b)} g$$

instead of the left-hand side. In this case, with $\int_{k(a)}^{k(b)} g$ given, we know $k(a)$ and $k(b)$ and wish to find a and b. That is, we want to *invert* the function k. This is possible if k is one–one. Writing α for $k(a)$ and β for $k(b)$ we then have the rule:

$$\int_{\alpha}^{\beta} g = \int_{h(\alpha)}^{h(\beta)} (g \circ k) \times Dk \qquad\qquad \textbf{Equation (7)}$$

where h is the inverse of the function k.

Example 2

Evaluate $\int_{-1}^{1} x \longmapsto \sqrt{1 + x}$ using $u = \sqrt{1 + x}$. Here we have:

$$\alpha = -1, \qquad \beta = 1$$
$$g(x) = \sqrt{1 + x} \qquad (x \in [-1, 1])$$
$$u = \sqrt{1 + x} \qquad (x \in [-1, 1]).$$

Notice that in this "backwards" form of the rule, it is h (the inverse of k) that maps x to u; so we choose:

$$h(x) = u = \sqrt{1 + x} \qquad (x \in [-1, 1])$$

which gives, on inverting this function,*

$$k(u) = u^2 - 1 \qquad\qquad (u \in [0, \sqrt{2}]).$$

Substitution in Equation (7) gives:

$$\int_{-1}^{1} x \longmapsto \sqrt{1 + x} = \int_{h(-1)}^{h(1)} (u \longmapsto g(u^2 - 1)) \times Dk(u)$$

$$= \int_{0}^{\sqrt{2}} (u \longmapsto u) \times 2u$$

$$= [\tfrac{2}{3}u^3]_0^{\sqrt{2}}$$

$$= \tfrac{4}{3}\sqrt{2}.$$

---

* We have

$$(h(x))^2 = 1 + x$$

so

$$x = (h(x))^2 - 1.$$

Exercise 3

Evaluate $\displaystyle\int_0^1 x \longmapsto x\sqrt{1-x^2}$ using Equation (7), with $u = \sqrt{1-x^2}$ (that is, $h(x) = \sqrt{1-x^2}$).

3.3   Additional Exercises

Exercise 1

Apply the rule of integration by parts to the integral $\displaystyle\int_a^b x\cos x$ treated in the text, taking $f(x) = \cos x$ and $Dg(x) = x$. Does the rule, applied in this way, help you to evaluate the integral? What lesson do you learn from this exercise?

Exercise 2

Evaluate $\displaystyle\int_0^\pi x \longmapsto \sin(3x)$ using the rule for integration by substitution with $k(x) = 3x$ $(x \in R)$.

3.4   Answers to Exercises

Section 3.1

Exercise 1

To evaluate

$$\int_0^\pi x \longmapsto x\sin x,$$

let

$$f(x) = x, \qquad\qquad \text{so that}\quad Df(x) = 1;$$
$$g(x) = -\cos x, \quad \text{so that}\quad Dg(x) = \sin x;$$
$$a = 0, \qquad b = \pi.$$

The formula:

$$\int_a^b f(x) \times Dg(x) = [f(x) \times g(x)]_a^b - \int_a^b g(x) \times Df(x)$$

becomes

$$\int_0^\pi x \sin x = [-x \cos x]_0^\pi - \int_0^\pi -\cos x$$

$$= -\pi \cos \pi + 0 \cos 0 + \int_0^\pi \cos x$$

$$= -\pi \times (-1) + 0 + [\sin x]_0^\pi$$

$$= \pi.$$

Exercise 2

Let

$$f(x) = x^2, \qquad \text{so that} \quad Df(x) = 2x,$$

$$g(x) = \exp x, \quad \text{so that} \quad Dg(x) = \exp x;$$

then we obtain:

$$\int x^2 \exp x = x^2 \exp x - \int \exp x \times 2x.$$

On page 70 we found that:

$$\int x \exp x = (x - 1) \exp x.$$

Combining these results we obtain:

$$\int x^2 \exp x = (x^2 - 2x + 2) \exp x.$$

That is, one of the primitive functions of $x \longmapsto x^2 \exp x$ is

$$x \longmapsto (x^2 - 2x + 2) \exp x.$$

Section 3.2

Exercise 1

(i) $DF(x) = \dfrac{1}{2\sqrt{x}} DG(\sqrt{x})$.

(ii) $\sin \sqrt{x} = \frac{1}{2} DG(\sqrt{x})$.

(iii) $DG(u) = 2 \sin u$.

(iv) $u \longmapsto -2 \cos u$.

(v) $G(u) = -2 \cos u + c$, where c is any constant.

(vi) $F(x) = -2 \cos \sqrt{x} + c$.

(vii) $[x \longmapsto -2\cos\sqrt{x}]_{\pi^2/4}^{\pi^2} = -2\cos\pi + 2\cos\dfrac{\pi}{2}$

$$= 2.$$

Exercise 2

Let F be a primitive function of $x \longmapsto x\exp(-x^2)$; as suggested, we guess that it has the form

$$F(x) = G(-x^2) \qquad (x \in [a, b]).$$

Then

$$DF(x) = -2xDG(-x^2) \qquad (x \in [a, b])$$

that is,

$$x\exp(-x^2) = -2xDG(-x^2) \qquad (x \in [a, b]),$$

or

$$DG(u) = -\tfrac{1}{2}\exp(u) \qquad (u \in [a^2, b^2]),$$

where

$$u = -x^2.$$

Now a primitive function for $u \longmapsto -\tfrac{1}{2}\exp(u)$ is

$$u \longmapsto -\tfrac{1}{2}\exp(u) \qquad (u \in [a^2, b^2]),$$

so we can take

$$F(x) = -\tfrac{1}{2}\exp(-x^2) \qquad (x \in [a, b]),$$

and find

$$\int_a^b x \longmapsto x\exp(-x^2) = -\tfrac{1}{2}\exp(-b^2) + \tfrac{1}{2}\exp(-a^2).$$

Exercise 3

Here, $u = h(x) = \sqrt{1 - x^2}$ $(x \in [0, 1])$, and h is a one–one function for this domain, so that it has an inverse given by

$$x = k(u) = \sqrt{1 - u^2} \qquad (u \in [0, 1]),$$

where $h(1) = 0$ and $h(0) = 1$.

Thus:

$$\int_0^1 x \longmapsto x\sqrt{(1-x^2)}$$

$$= \int_1^0 \left(u \longmapsto \sqrt{1-u^2} \times \sqrt{1-(1-u^2)}\right) \times \frac{-u}{\sqrt{1-u^2}}$$

$$= \int_1^0 u \longmapsto -u^2$$

$$= \left[-\tfrac{1}{3}u^3\right]_1^0$$

$$= \tfrac{1}{3}.$$

Section 3.3

Exercise 1

Let

$$f(x) = \cos x, \quad \text{so that} \quad Df(x) = -\sin x,$$
$$g(x) = \tfrac{1}{2}x^2, \quad \text{so that} \quad Dg(x) = x;$$

then we obtain:

$$\int_a^b x \cos x$$

$$= \left[\tfrac{1}{2}x^2 \cos x\right]_a^b - \int_a^b -\tfrac{1}{2}x^2 \sin x.$$

This time the new integral is *more complicated* than the one we started with (the power of x in it is higher). The lesson to be learnt is that if there are several possible ways of choosing f and g in the formula for integration by parts, it is worth trying all of them if you do not at first find one that simplifies the integral. (A more advanced lesson might be that if the function to be integrated is a polynomial function times another function, it is better to make the polynomial f rather than Dg, because differentiating a polynomial reduces the degree while integrating increases it.)

Exercise 2

Applying the rule with $a = 0$ and $b = \pi$, we obtain:

$$\int_0^\pi (g \circ k) \times Dk = \int_{k(0)}^{k(\pi)} g.$$

The integrand is

$$g(k(x)) \times Dk(x) = \sin(3x),$$

and we have $Dk(x) = 3$; so we want

$$3g(k(x)) = \sin(3x),$$

that is, $g(u) = \frac{1}{3}\sin u$, where $u = 3x$.
The given integral is therefore equal to

$$\int_{k(0)}^{k(\pi)} \frac{1}{3}\sin u = [-\frac{1}{3}\cos u]_0^{3\pi}$$

$$= -\frac{1}{3}\cos 3\pi + \frac{1}{3}\cos 0$$

$$= \frac{1}{3} + \frac{1}{3}$$

$$= \frac{2}{3}.$$

CHAPTER 4   SOME APPLICATIONS OF INTEGRATION

4.0   Introduction

In this chapter we introduce a number of applications of integration. These applications will use principles discussed in Chapters 7 and 9 of Volume 1 and techniques which were introduced in the previous Chapter.

In Volume 1, Chapter 7 we defined the definite integral of f in $[a, b]$ as the limit of the sequence

$$S_1, S_2, S_3, \ldots, S_n, \ldots$$

where

$$S_n = h[f(a) + f(a + h) + \cdots + f(a + \{n - 1\}h)]$$

We have also shown that provided the images are positive, this integral will give us the area beneath the graph of f between a and b. In this chapter we intend to look at other physical situations which, when we analyze them, produce sums of terms which we can identify with a definite integral. For example, the terms in the sequence $S_1, S_2, S_3, \ldots S_n, \ldots$, which were used to represent the sums of the areas of rectangles, can be used to represent the sums of volumes of standard shapes.

We start therefore with *volumes of solids of revolution* and then proceed to a brief discussion of *averages* and of *velocity* and *distance*.

In practical cases, we often find that our analytic techniques of integration become impracticable and we have to resort to approximation methods. We discuss two simple methods of approximation before proceeding to two further practical applications of techniques introduced earlier in Chapter 3.

4.1   Volume of a Solid of Revolution

Consider a region bounded by the graph of a function f (whose images are positive in $[a, b]$), the lines specified by $x = a$ and $x = b$, and the interval of the x-axis between a and b. Focus attention on the boundary of such a region, and imagine the boundary being rotated round the x-axis; so that it generates the bounding surface of a solid. We usually refer to the rotation of the graph of f in $[a, b]$ (leaving the rotation of the area to be understood.) The volume of such a solid is called a *volume of*

revolution, and the calculation of these volumes is simply an application of the definite integral.

As an illustration, we shall investigate the problem of finding the volume of a cone generated by rotating the hypotenuse of a right-angled triangle around the x-axis.
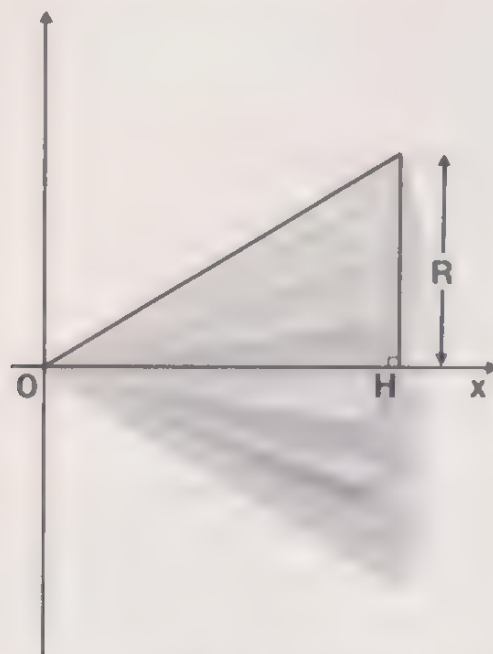
You may know that the volume, V, is given by the formula $V = \frac{1}{3}\pi R^2 H$ where H is the height of the cone and R is the radius of the base. We shall derive this formula.

To correspond to our elementary rectangles (used in the calculation of area) we choose discs of thickness h. They are similar to the discs on the child's toy shown in the diagram.
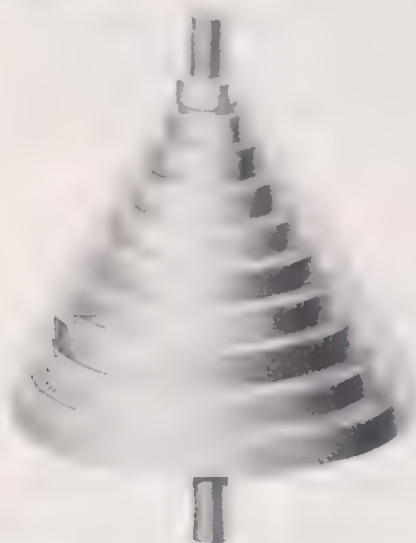
If there are n discs we have

$$h = \frac{H}{n},$$

and the edge of the cone will be the graph of the function

$$x \longmapsto \frac{Rx}{H} \qquad (x \in [0, H])$$
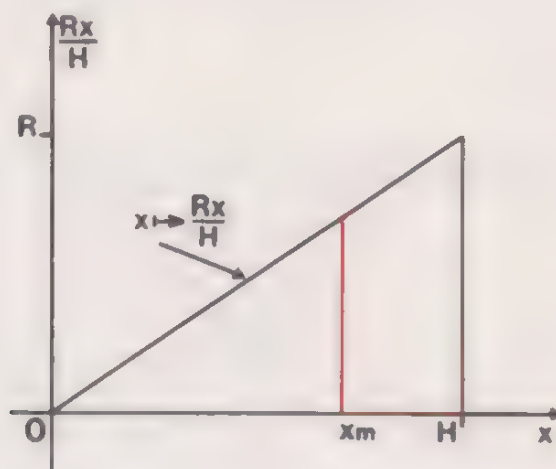
The radius of each of the n discs will be given by an ordinate. Let the n ordinates be at $x_0, x_1, \ldots, x_{n-1}$, then

$$x_m = mh \qquad (m = 0, 1, 2, \ldots, n - 1)$$

and the corresponding ordinate is the image of x_m, so that

$$\frac{Rx_m}{H} = \frac{Rmh}{H}$$

and the volume of the elementary disc is therefore

$$\pi \left(\frac{Rmh}{H} \right)^2 h$$

The total volume of these n discs is

$$V_n = 0 + \frac{\pi R^2 h^3}{H^2} \times 1^2 + \cdots + \frac{\pi R^2 h^3}{H^2} \times m^2 + \cdots$$

$$+ \frac{\pi R^2 h^3}{H^2} \times (n - 1)^2$$

$$= h \left[0 + \frac{\pi R^2}{H^2} h^2 + \cdots + \frac{\pi R^2}{H^2} (mh)^2 + \cdots + \frac{\pi R^2}{H^2} \{(n - 1)h\}^2 \right]$$

We write the expression for V_n this way because we are going to compare it with the sum whose limit we know to be a definite integral, rather than go through the lengthy algebraic process of finding the sum again. Thus, comparing V_n with S_n, given on page 83:

$$S_n = h[f(a) + f(a + h) + \cdots + f(a + mh) + \cdots$$
$$+ f(a + \{n - 1\}h)]$$

it is fairly easy to see that the appropriate function is

$$f : x \longmapsto \frac{\pi R^2}{H^2} x^2$$

and that the end-points of integration are $a = 0$ and $b = nh + a = nh = H$.
Thus

$$\lim V = \lim S = \int_0^H x \longmapsto \frac{\pi R^2 x^2}{H^2}$$
$$= \tfrac{1}{3}\pi R^2 H$$

Old hands at the game would abbreviate the above comparison even
further. They would jump straight from the volume of the elementary
disc in the mth position:

$$\pi \left(\frac{Rx_m}{H}\right)^2 h$$

to the appropriate function and then to the integral, by dropping the
"h" and the subscript "m". In fact, the justification for doing this is
outlined by the fuller argument above.

We can generalize the definite integral for finding volumes of revolution.
If the volume of revolution is formed by the graph of f rotating about the
x-axis between $x = a$ and $x = b$, then the volume generated is

$$\int_a^b x \longmapsto \pi\{f(x)\}^2 \quad \text{or} \quad \pi \int_a^b x \longmapsto \{f(x)\}^2$$

(We could also use $\pi \int_a^b f^2$ but f^2, which here stands for $f \times f$, could be
confused with $f \circ f$.)

In the alternative Leibniz notation we would have

$$\int_a^b \pi\{f(x)\}^2 \, dx \quad \text{or} \quad \pi \int_a^b y^2 \, dx$$

where $y = f(x)$.

Exercise 1
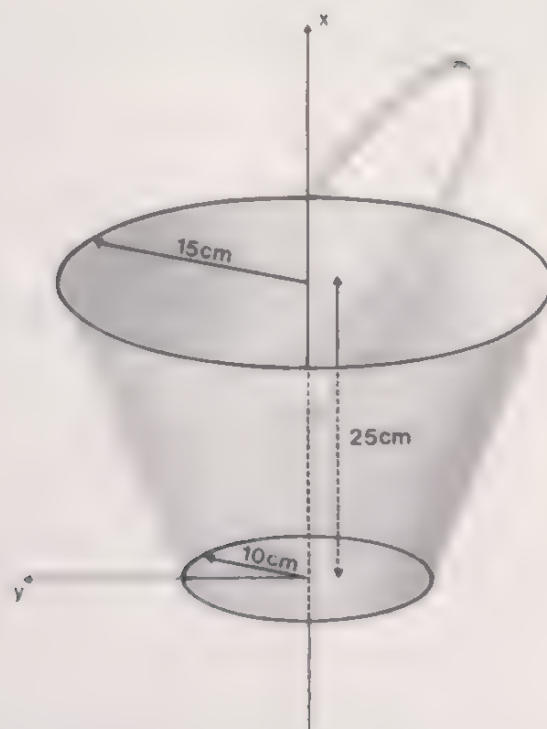
Find the volume of a bucket of circular cross-section with the dimensions shown.

4.2 Averages

When faced with an excavation problem, a builder might wish to estimate in advance the total volume of earth to be removed. One way of estimating such a volume involves determining a cross-sectional area using given measurements. In actual practice, when viewing a site for building operations, a builder will often mentally "add a bit here" and "take a bit off there" and take an average depth to estimate the cross-sectional area. For example, faced with the site illustrated below a builder might assume that the area of the red region was the same as the area of the black region, and that the cross-sectional area required was the same as the area of the rectangle *OABC*.
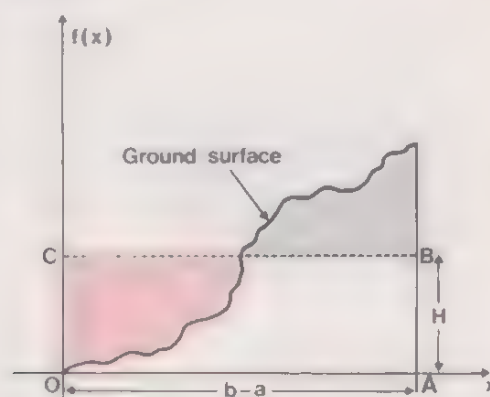
He would probably call the height of this rectangle "the *average* height" of the site. If we suppose that the curve is the graph of a function f between a and b relative to an appropriate axis, then the cross-sectional area is, of course,

$$\int_a^b f$$

The base of our rectangle is then of length $b - a$, and if its height is H, we have

$$\text{average of } f(x) \text{ over } [a, b] = H = \frac{1}{b - a} \int_a^b f$$

which we can abbreviate to:

$$\text{average} = \frac{1}{b - a} \int_a^b f$$

(We used this formula in our discussion of the fundamental theorem of the calculus, Volume 1, Chapter 9.)

Exercise 1

Find the average of $f(x)$ over $[0, 4]$, where f is the function:

$$x \longmapsto x^2 \qquad (x \in [0, 4])$$

4.3   Velocity and Distance

A hiker goes on a walk lasting four hours, excluding rests. From past experience he reckons that he walks at an average speed of 5 km/h. You conclude that his walk was approximately 20 km long. This was all straightforward since
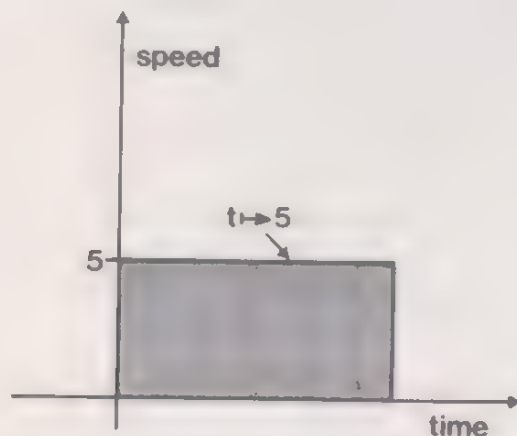
speed × time = distance

you assumed that the speed was constant, and so the calculation was simple. By plotting speed against time on a graph, we see that the area beneath the graph (the area of a rectangle in this case) represents the distance covered.

Similarly, if the speed is not constant, the distance covered is also represented by the area under an appropriate graph. We can see that this is so by dividing the time interval into n equal sub-intervals and then assuming that the speed is constant over each of these time-sub-intervals. The distance travelled during each sub-interval is then represented by the area of a rectangle, and the total distance covered is the sum of the areas of the n rectangles; this will be a close approximation to the area under the graph when n is large.

Here we prefer to use "velocity" rather than "speed" because velocity means speed in a known direction. When we consider motion in a directed straight line, we take velocity to be positive if its direction is the same as that of the line, and negative if it has the opposite direction.

The calculation of the distance travelled is not so easy as it was for the case of the hiker because the required area is given by a definite integral. We must be careful of the physical interpretation of "distance" in relation to the definite integral when the value of the velocity function becomes negative. In any particular context does it mean the *total* distance travelled or the distance that the object is from its starting point? This difference of meaning is illustrated in the following exercise.

Exercise 1

A ball, thrown vertically upwards with an initial velocity of 20 m/s, has a velocity at time t seconds given approximately by

$$v(t) = (20 - 10t)\,\text{m/s}$$

Determine

(i) the number of metres the ball is above the ground after 3 seconds;
(ii) the number of metres that it has travelled in that time.

4.4 Approximation Methods

In Volume 1, Chapter 7 we defined the definite integral of a function as the limit of a sequence, and we have shown that for certain cases (simple polynomial functions) we can evaluate this limit. In Volume 1, Chapter 9

and Volume 2, Chapter 3 we obtained general results which allowed us to extend the set of functions for which this evaluation is practicable. But even so there are still many functions for which that general process is impracticable, for example,

$$\int_0^1 x \longmapsto \frac{1}{\sqrt{x^3 + 1}},$$

or for which it is unnecessarily complicated, for example,

$$\int_0^{0.5} x \longmapsto \frac{x^5}{\sqrt{1 - x^2}}$$

When we meet integrals such as these in practical work, we usually need a numerical answer correct to some given accuracy, so we can often usefully return to our original approximation processes (or variations of these). Thus we obtain an *estimate* of the limit to a given accuracy, rather than an exact formula for the limit.

The Trapezoidal Rule

Suppose that we wish to find the area under part of a parabola. A good estimate can be found by using

$$\tfrac{1}{2}(A_n + a_n)$$

where A_n = the sum of the n areas of the larger rectangles, and a_n = the sum of the n areas of the smaller rectangles where n = the number of intervals.[*]
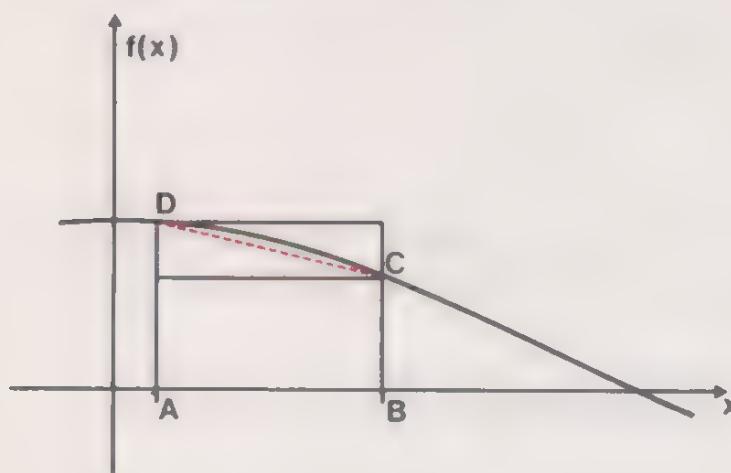
* See Volume 1, Chapter 7, Example 1.

What does this mean geometrically? Half the sum of one larger rectangle and one smaller rectangle (an example is outlined in the figure) equals the area of the trapezium $ABCD$ (with upper boundary, CD, shown by the dashed line in the diagram). Thus the approximation is equivalent to drawing the set of dashed lines (one for each interval) as the upper boundary to the area. That is, we approximate to the total area by the sum of the areas of the trapezia thus constructed.

We now turn to the case where $f(x)$ is non-negative in $[a, b]$.

Suppose we wish to find $\int_a^b f$ where the graph of f is given below:
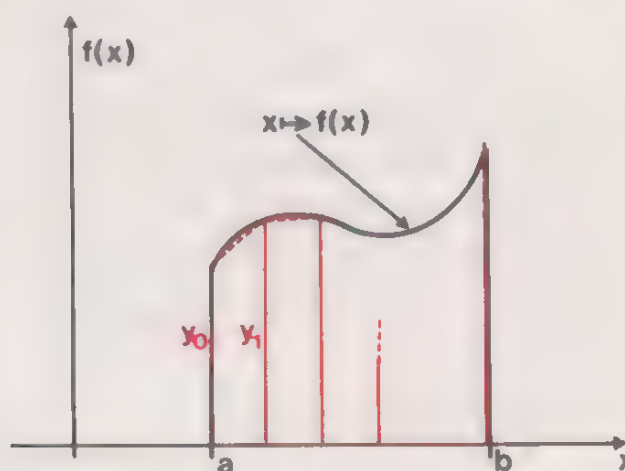
We construct the set of dashed lines just as in the case of the parabola. Suppose the ordinates of points on the graph at

$$a, a + h, a + 2h, \ldots, a + nh = b$$

are

$$y_0, y_1, y_2, \ldots, y_n \quad \text{respectively}$$

where

$$h = \frac{b - a}{n}$$

We have:

(area of first trapezium on left) $= \frac{1}{2}(y_0 + y_1) \times h$

(area of second trapezium from left) $= \frac{1}{2}(y_1 + y_2) \times h$

and so on, until

(area of last trapezium) $= \frac{1}{2}(y_{n-1} + y_n) \times h$

Therefore, adding these equations together, we get:

the total area of all the trapezia

$$= \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n)$$

and therefore

$$\int_a^b f \simeq \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n)$$

where $\simeq$ means "is approximately equal to".

This is known as the trapezoidal rule for evaluating definite integrals.*

We must now ask the question: "How approximate is 'approximate'?"

To answer this question we again consider the parabola. We use the fact that, for the rectangle approximation, provided the graph of the function slopes downwards over the whole interval or upwards over the whole interval, the difference, $A_n - a_n$, is the area of the largest rectangle. (*See* Volume 1, Chapter 7.) In the case of the parabola:

$$A_n - a_n = h|f(b) - f(a)| \quad \text{where } h = \frac{b-a}{n}$$
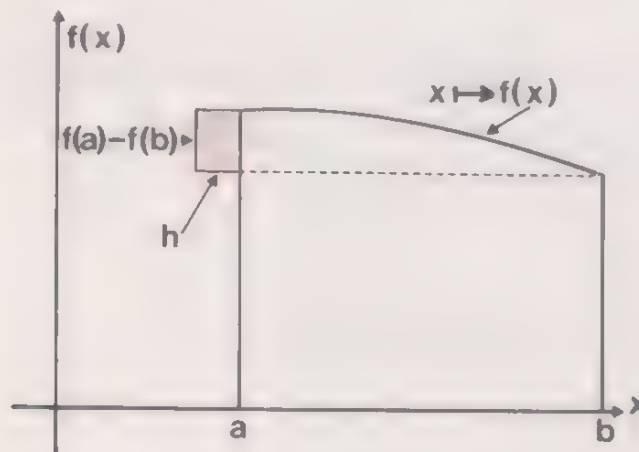
The maximum possible error in the best estimate of the area, $\frac{1}{2}(A_n + a_n)$, is $\frac{1}{2}(A_n - a_n)$ which is

$$\frac{h}{2}|f(b) - f(a)| \quad \text{in the case of the parabola.}$$

The thing to notice about this formula is that, since $f(b)$ and $f(a)$ do not depend on the number of rectangles we choose to take in our approxi-

* See note on page 97

mation, the maximum error is proportional to h, the width of each sub-interval.

We return to the general case. The difference between this case and that of the parabola is that the graph of the function f no longer slopes downwards over the whole interval or upwards over the whole interval. (See the second figure on page 91.) However, we have only to split $[a, b]$ into smaller intervals in each of which the slope has the same sign throughout, and to apply the above argument to each of the smaller intervals. Since the "rectangle" method is equivalent to our trapezoidal rule, this means that the best we can say at this stage about the maximum error for the trapezoidal rule is that it too is proportional to h.*

In fact, it is usually more accurate than the rectangle method.

Example 1

How many intervals would you need to take, *at most*, to evaluate

$$\int_0^1 x \longmapsto \frac{1}{x^3 + 1}$$

to an accuracy of two decimal places using the trapezoidal rule?

In practice, when evaluating $\frac{1}{x^3 + 1}$, we would round off the image values to some convenient number of decimal places, so that the maximum error on each image will be ε, say. To how many decimal places must the images be calculated to ensure that the integral is accurate to two decimal places, and how many intervals will then be needed to ensure that the overall accuracy is to two decimal places?

Using the trapezoidal rule gives a maximum error of

$$\frac{h}{2}|f(1) - f(0)| = \frac{h}{2}|\tfrac{1}{2} - 1| = \frac{h}{4}$$

where

$$f : x \longmapsto \frac{1}{x^3 + 1} \qquad b = 1 \text{ and } a = 0$$

($f(x)$ decreases as x increases in $[0, 1]$.)

---

* It can be shown that the maximum error is much smaller than suggested here; it is proportional to h^2

For accuracy to 2 decimal places the maximum error must be less than or equal to $0.005 = 5 \times 10^{-3}$.

Therefore

$$\frac{h}{4} \leqslant 5 \times 10^{-3}$$

so

$$h \leqslant 2 \times 10^{-2}$$

and then, if the number of intervals is n, we have

$$n = \frac{b-a}{h} = \frac{1}{h} \geqslant 50$$

Therefore a minimum of 50 intervals is required to *guarantee* the required accuracy (with our present knowledge about the accuracy of the trapezoidal rule).

We have

$$\int_a^b f \simeq \frac{h}{2} \{y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n\}$$

On the right-hand side there are $2n$ values of $f(x)$ inside the brackets. If each ordinate has an inherent error ε, the total error on the right-hand side will be

$$\frac{h}{2} \times 2n\varepsilon = nh\varepsilon = (b-a)\varepsilon$$

that is, a total error of ε in this particular case because $(b-a) = 1$ (note that this is independent of the number of intervals).

To be really sure that our *total* error from the use of the trapezoidal rule *and* the inexact data does not exceed 0.005, we can use 100 intervals (the error introduced by the trapezoidal rule from this is then $\leqslant 0.0025$) together with data accurate to 3 decimal places (error from this is then $\leqslant 0.0005$) to give a total possible error of 0.0030; but in fact 50 intervals and data accurate to 3 decimal places would almost certainly suffice.

Exercise 1

Repeat the last example with the definite integral

$$\int_0^2 x \longmapsto \exp(-x^2)$$

with a required accuracy of 3 decimal places. $[e^{-4} = 0.0183]$.

Simpson's Rule

The trapezoidal rule gave an approximation to the area under a curve by using a set of straight line boundaries. One obvious way to improve the quality of the approximation is to take some account of the curvature of the boundary. In deriving the trapezoidal rule we took each pair of consecutive points $(a + mh, y_m)$ and $(a + (m + 1)h, y_{m+1})$, $m = 0, \ldots, n - 1$, found the straight line which passed through them (although we did not specify its equation since we already knew the area of a trapezium) and used this as the upper boundary of the area. We now introduce (no more than that) Simpson's rule for the area under a curve.

The basic element of area is now the one shown in the diagram covering two intervals with a parabolic upper bounding surface. Thus we now need to split the total interval into "2-interval" or "3-point" subsets.
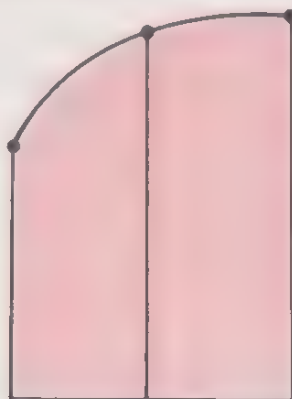
What implication does this have on the number of sub-intervals we use? The answer to this is that the number of sub-intervals we use must now be even.

The bare outline of the basic steps in the argument are continued in the following text. You may like to check some of the steps and fill in the detail

Consider a "3-point" subset in which the 3 points to be fitted on the given curve have co-ordinates $(-h, y_0)$, $(0, y_1)$, (h, y_2). The approximating quadratic polynomial function has the form

$$f : x \longmapsto a_2 x^2 + a_1 x + a_0 \qquad (x \in [-h, h])$$

It can be shown:

(i) by evaluating $\int_{-h}^{h} f$, that the area beneath the graph of f is

$$\frac{2a_2 h^3}{3} + 2a_0 h$$

(ii) by solving three simultaneous equations that

$$a_0 = y_1$$

$$a_1 = \frac{y_2 - y_0}{2h}$$

$$a_2 = \frac{y_0 - 2y_1 + y_2}{2h^2}$$

(iii) by substituting the appropriate results in (ii) into the result in (i) that the area beneath the graph of f is

$$\frac{h}{3}(y_0 + 4y_1 + y_2)$$

This result gives the approximation to the area beneath the original given curve in any "3-point" interval in which the ordinates are y_0, y_1 and y_2. Thus in the next "3-point" interval with ordinates y_2, y_3, y_4 (the interval in the figure shown dotted in black and red), the area approximation is

$$\frac{h}{3}(y_2 + 4y_3 + y_4)$$
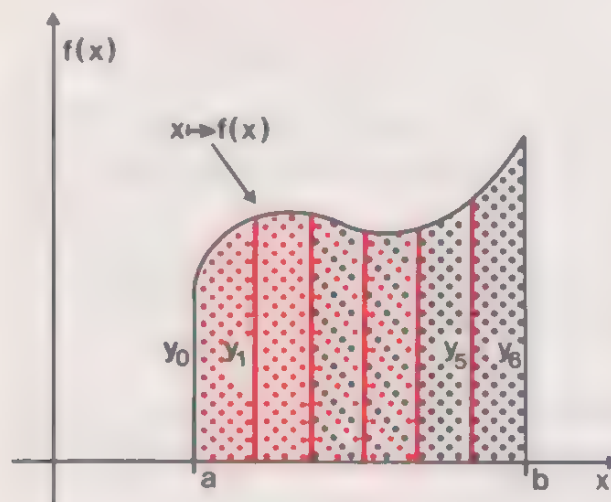
Continuing in this way, we obtain the approximation (in the case of the function illustrated) to the total area as

$$\frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + 4y_5 + y_6)$$

In fact, the general formula with n intervals (n even) is

$$\int_a^b f \simeq \frac{h}{3}(y_0 + 4y_1 + 2y_2 + \cdots + 4y_{n-1} + y_n)$$

This formula is known as Simpson's rule.

It is generally true that you can obtain a better approximation (with the same number of intervals) using Simpson's rule than with the trapezoidal rule, but again we need calculus to demonstrate this more explicitly.

Note

We have obtained Simpson's rule and the trapezoidal rule for $\int_a^b f$ by considering the area beneath the graph of f between a and b. For simplicity, we have considered the special case in which $f(x) > 0$ for all x in $[a, b]$. In fact, these rules apply when $f(x)$ is not always positive in $[a, b]$. This can be seen by again considering areas (taking care of the signs) and slightly modifying our derivations.

4.5 An Application of Integration by Parts

It is found experimentally that the way in which a radioactive substance, such as uranium, decays is described to a very good approximation by the formula

$$N(t) = A \exp(-ct) \qquad (t \in R^+) \tag{Equation (1)}$$

where A is a positive number, c is a positive number called the *decay constant*, and N is the function defined by

$$N : \begin{pmatrix} \text{time, } t, \text{ measured} \\ \text{in years, say,} \\ \text{since some} \\ \text{arbitrary initial} \\ \text{instant} \end{pmatrix} \longmapsto \begin{pmatrix} \text{number of uranium} \\ \text{atoms remaining} \\ \text{at time } t \end{pmatrix} \qquad (t \in R^+).$$
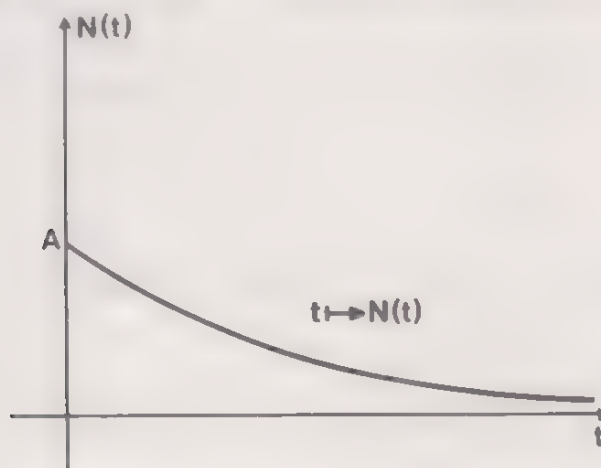
The problem is to find the *mean life-time* of the uranium atoms; that is to say, the average time a uranium atom lasts before decaying. This average can be expressed as a definite integral, which we shall evaluate using integration by parts.

The average life-time of the atoms is defined by the equation:

$$\text{average life-time} = \frac{\text{sum of the life-times of all the atoms}}{\text{number of atoms}}$$

Using the techniques developed earlier, we can approximate to the numerator by an integral over the time interval $[0, T]$, where T is some very large number. We divide the time interval $[0, T]$ into m equal sub-intervals, $[0, t_1]$, $[t_1, t_2]$, $[t_2, t_3]$, . . . , $[t_{m-1}, T]$, where m is any positive integer.
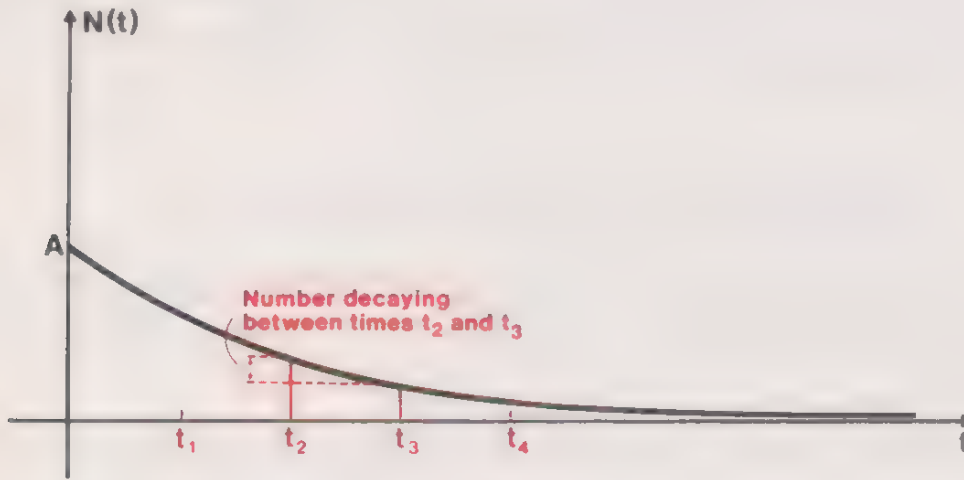
The length of each sub-interval is $\frac{T}{m}$. Consider any one of these sub-intervals, $[t_{k-1}, t_k]$, say. Then the number of atoms whose times of decay lie in the interval $[t_{k-1}, t_k]$ is $N(t_{k-1}) - N(t_k)$.

Provided m is large, so that the interval length $\frac{T}{m} = t_k - t_{k-1}$ is small, we can make the approximation that all these atoms decay at the end of the interval, that is, at the instant t_k, so that their contribution to the sum of the life-times of all atoms decaying in the interval $[0, T]$ is approximately

$$(N(t_{k-1}) - N(t_k))t_k,$$

and this sum of the life-times itself is given approximately by:

$$\left\{ \begin{array}{l} \text{sum of the life-times} \\ \text{of atoms decaying} \\ \text{during } [0, T] \end{array} \right\} \simeq \sum_{k=1}^{m} (N(t_{k-1}) - N(t_k))t_k.$$

<div style="text-align: right;">Equation (2)</div>

To use the definition of a definite integral of a function f in the form:

$$\int_0^T f = \lim_{m \text{ large}} \frac{T}{m} \sum_{k=1}^{m} f(t_k)$$

we would like to approximate the factor $N(t_{k-1}) - N(t_k)$ in Equation (2) by something that depends only on t_k, and is proportional to $\dfrac{1}{m}$. Now the factor we wish to approximate, representing the number of atoms decaying during the time interval $[t_{k-1}, t_k]$, is equal to the interval length $\dfrac{T}{m}$ multiplied by the average rate of decrease of $N(t)$ during the interval.

the slope at this point $=$ DN(t_{k-1})

this slope $=$ average rate of decrease of N(t) over the interval $[t_{k-1}, t_k]$

decrease

$\frac{T}{m}$

t_{k-1} t_k t

So Equation (2) becomes

$$\left\{ \begin{array}{l} \text{sum of the life-times} \\ \text{of atoms decaying} \\ \text{during } [0, T] \end{array} \right\} = \frac{T}{m} \sum_{k=1}^{m} \left(\frac{N(t_{k-1}) - N(t_k)}{t_k - t_{k-1}} \right) t_k.$$

Equation (3)

The point of this manipulation is that, for very small interval length, we can approximate the average rate of change by the local rate of change at some point in the interval, say the end-point, t_k. This local rate of change is the *derivative* at t_k, and so Equation (3) gives the further approximation:

$$\left\{ \begin{array}{l} \text{sum of the life-times} \\ \text{of atoms decaying} \\ \text{during } [0, T] \end{array} \right\} \simeq \frac{T}{m} \sum_{k=1}^{m} (-t_k DN(t_k)).$$

Equation (4)

where DN denotes the derived function of N. (Since $N(t)$ is decreasing, $-DN(t)$ is positive.)

Taking the limit for very large m, the approximation that all the atoms decay at time t_k, and the approximation of replacing an average slope

by a local slope, both become exact, and so Equation (4) becomes

$$\left\{ \begin{array}{l} \text{sum of the life-times} \\ \text{of atoms decaying} \\ \text{during } [0, T] \end{array} \right\} = \int_0^T t \longmapsto -t \, DN(t).$$

When T is very large, nearly all the atoms present at time 0 will have decayed during $[0, T]$, and so we expect the contribution from the rest of the atoms to the total life to be very small. By taking the limit, we can justify this expectation; the phrases "nearly all" and "very small" in the preceding sentence then become "all" and "zero" respectively, and we have:

$$\left\{ \begin{array}{l} \text{sum of the life-times} \\ \text{of all the atoms} \end{array} \right\} = \lim_{T \text{ large}} \int_0^T t \longmapsto -t \, DN(t).$$

Dividing this by the total number of atoms that decay, which is A, we obtain the formula:

$$\left\{ \begin{array}{l} \text{average} \\ \text{life-time} \end{array} \right\} = \lim_{T \text{ large}} \frac{1}{A} \int_0^T t \longmapsto -t \, DN(t). \qquad \text{Equation (5)}$$

This completes the difficult part of the problem, which is the conversion of the real situation into a mathematical formula. All that remains is to evaluate the integral for the given function N.

The given function N is defined by $\qquad N(t) = A \exp(-ct) \qquad (t \in R^+)$,

so $\qquad DN(t) = -cA \exp(-ct)$,

and substituting this into Equation (5) we find

$$\left\{ \begin{array}{l} \text{average} \\ \text{life-time} \end{array} \right\} = \lim_{T \text{ large}} \int_0^T t \longmapsto ct \exp(-ct). \qquad \text{Equation (6)}$$

The formula for integration by parts from Chapter 3, Section 1 is

$$\int_a^b t \longmapsto f(t) Dg(t) = [t \longmapsto f(t)g(t)]_a^b - \int_a^b t \longmapsto g(t) Df(t)$$

and here we shall take

$$f(t) = ct \qquad Dg(t) = \exp(-ct)$$
$$Df(t) = c \qquad g = ?$$

For g we need a function whose derived function is $t \longmapsto \exp(-ct)$. We know that the exponential function is its own derived function; this suggests trying $t \longmapsto \exp(-ct)$ for g. In fact the derived function of $t \longmapsto \exp(-ct)$ is $t \longmapsto -c \exp(-ct)$, which is not quite the Dg that

we want; but we can remove the unwanted factor $(-c)$ by taking

$$g : t \longmapsto \frac{\exp(-ct)}{c}$$

instead. Substituting for $f(t)$ and $g(t)$ in the integration-by-parts formula, with $a = 0$ and $b = T$, we find:

$$\int_0^T ct \exp(-ct) = \left[ct \frac{\exp(-ct)}{-c} \right]_0^T - \int_0^T \frac{\exp(-ct)}{-c} c$$

$$= [-t \exp(-ct)]_0^T - \frac{1}{c} [\exp(-ct)]_0^T$$

$$= -T \exp(-cT) + 0 - \frac{\exp(-cT)}{c} + \frac{1}{c}$$

so that, by Equation (6)

$$\left\{ \begin{array}{l} \text{average} \\ \text{life-time} \end{array} \right\} = \frac{1}{c} + \lim_{T \text{ large}} \left\{ -T \exp(-cT) - \frac{1}{c} \exp(-cT) \right\}.$$

Equation (7)

This disposes of the integration.

The last step is to deal with the limit in Equation (7).

The graph of the function N shows that the term $\frac{1}{c} \exp(-cT)$ has limit 0 for large t. The limit of the other term is not quite so obvious, because the small quantity $\exp(-cT)$ is multiplied by a factor T which is large, so that it is not immediately clear whether their product is large or small. Calculation shows, however, that the product is very small for large T, as can be seen from the table below; and in fact it is possible to prove (for $c > 0$) that $\lim_{T \text{ large}} (cT \exp(-cT)) = 0.$

$x (= cT)$	$x \exp(-x)$
0	0
1	0.368
2	0.257
3	0.149
4	0.073
5	0.034
6	0.015
7	0.006
8	0.003
9	0.001
10	0.000

Thus each term in the limit in Equation (7) has limit zero, and the formula reduces to:

$$\text{average life-time} = \frac{1}{c}$$

This answers the problem posed at the beginning of this section.

Summary

1   Number of uranium atoms remaining at time t is given by

$$N(t) = A \exp(-ct) \qquad (t \in R^+).$$ **Equation (1)**

2   Average life-time of atoms is defined by

$$\text{average life-time} = \frac{\text{sum of the life-times of all the atoms}}{\text{number of atoms}}$$

3   Number of atoms decaying in time interval $[t_{k-1}, t_k]$ is

$$N(t_{k-1}) - N(t_k),$$

and if we assume that they all decay at the instant t_k, the total life-time of these atoms is, approximately,

$$(N(t_{k-1}) - N(t_k))t_k.$$

4   We divide the interval $[0, T]$, where T is large, into m equal sub-intervals. The total life time of all atoms decaying in this interval is, approximately,

$$\sum_{k=1}^{m} (N(t_{k-1}) - N(t_k))t_k$$

$$= \frac{T}{m} \sum_{k=1}^{m} \left(\frac{N(t_{k-1}) - N(t_k)}{t_k - t_{k-1}} \right) t_k, \quad \text{since } \frac{T}{m} = t_k - t_{k-1},$$

                                                                       Equations (2) and (3)

$$\simeq \frac{T}{m} \sum_{k=1}^{m} (-t_k DN(t_k)).$$ **Equation (4)**

5   In the limit, for very large m,

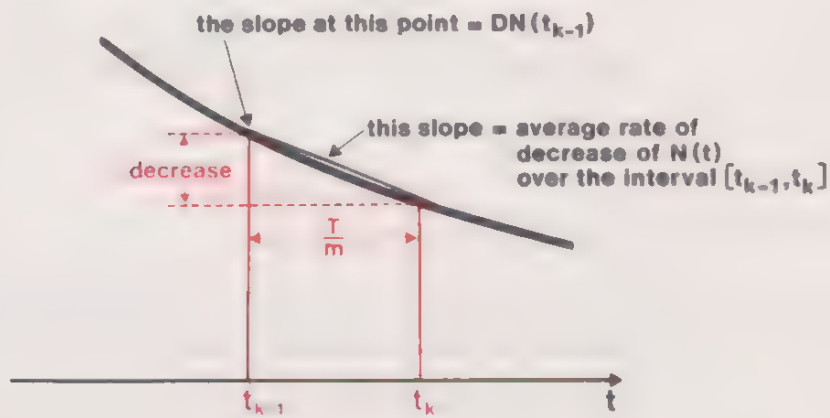$$\left\{ \begin{array}{l} \text{sum of life-times} \\ \text{of atoms decaying} \\ \text{during } [0, T] \end{array} \right\} = \int_0^T t \longmapsto -t \, DN(t)$$

and so

$$\begin{Bmatrix} \text{sum of the life-times} \\ \text{of all the atoms} \end{Bmatrix} = \lim_{T \text{ large}} \int_0^T t \longmapsto -t \, DN(t)$$

$$\begin{Bmatrix} \text{average} \\ \text{life-time} \end{Bmatrix} = \lim_{T \text{ large}} \frac{1}{A} \int_0^T t \longmapsto -t \, DN(t)$$

Equation (5)

$$= \lim_{T \text{ large}} \int_0^T t \longmapsto ct \exp(-ct)$$

Equation (6)

$$= \frac{1}{c} + \lim_{T \text{ large}} \left(-T \exp(-cT) \right.$$

$$\left. - \frac{1}{c} \exp(-cT) \right)$$

Equation (7)

$$= \frac{1}{c}.$$

4.6   An Application of Integration by Substitution

As an illustration of the rule of integration by substitution, let us apply it to the problem of calculating the area enclosed by an ellipse, the curve giving the shape of the orbit of a planet or comet moving under the gravitational influence of the sun.

For our purposes the ellipse may be defined as the graph of the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Equation (1)

where x and y are Cartesian co-ordinates, and a and b are positive real numbers.
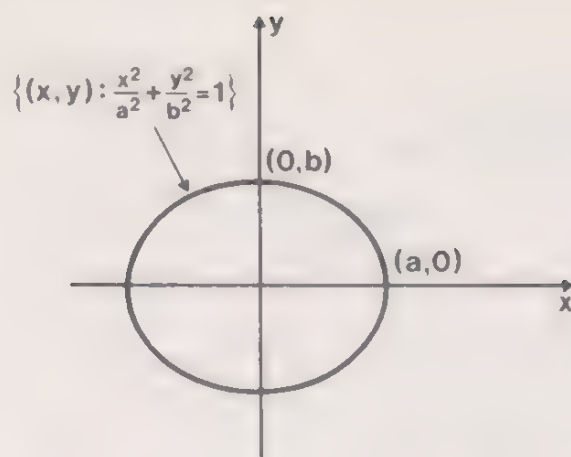
The co-ordinate axes cut the ellipse into four congruent parts, and the total area of the ellipse is just four times the area of any one of them, say the part for which $x > 0$ and $y > 0$.
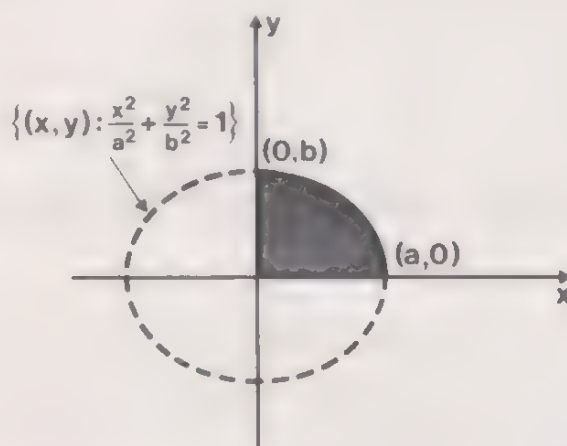
This quarter-ellipse is the type of area that we can express as an integral. The formula for the area is:

$$\text{area of quarter-ellipse} = \int_0^a f.$$

Here f is the function whose graph is the part of the ellipse shown in the figure. The domain of this function is $[0, a]$ and in view of Equation (1) it must satisfy

$$\frac{x^2}{a^2} + \frac{(f(x))^2}{b^2} = 1 \qquad\qquad (x \in [0, a]),$$

or, solving for $f(x)$,

$$f(x) = b \sqrt{1 - \frac{x^2}{a^2}} \qquad (x \in [0, a]).$$

(The negative square root $f(x) = -b\sqrt{1 - \dfrac{x^2}{a^2}}$ would also be a solution, but would give the wrong part of the ellipse – the part below the x-axis.) We now have:

$$\text{area of quarter-ellipse} = \int_0^a x \longmapsto b\sqrt{1 - \frac{x^2}{a^2}}.$$

The substitution (the choice of the function $h : x \longmapsto u$) which enables us to evaluate this integral is not as obvious as the ones we used in Chapter 3. However, the presence of the square root sign with the square of $\dfrac{x}{a}$ inside it suggests the use of Pythagoras' theorem, applied to the triangle shown below.
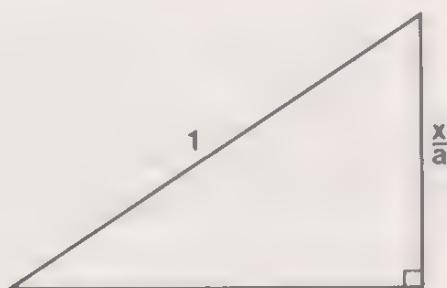
Pythagoras' theorem tells us that the base of this triangle is $\sqrt{1 - \dfrac{x^2}{a^2}}$, which is the unpleasant part of the integrand.
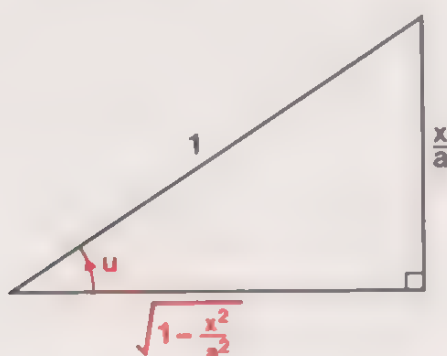
If we call the angle at the left of this triangle u then we have:

$$\sin u = \frac{x}{a},$$

$$\cos u = \sqrt{1 - \frac{x^2}{a^2}}.$$

so that there is a chance of simplifying the integral by the substitution of $a \sin u$ for x. The rule for integration by substitution, with the end-points altered to conform to the notation of this section, is:

$$\int_0^a x \longmapsto f(x) = \int_{h(0)}^{h(a)} (u \longmapsto f(k(u))) \times Dk(u) \qquad \text{Equation (2)}$$

where

$$h:x \longmapsto u \qquad (x \in [0, a]),$$

$$k:u \longmapsto x \qquad (u \in [h(0), h(a)]).$$

The substitution $\sin u = \dfrac{x}{a}$ corresponds to:

$$k(u) = a \sin u$$

$$h(x) = \text{the angle in the first quadrant, with } \sin (h(x)) = \frac{x}{a}.$$

The triangle shows that $h(x)$ increases with x for $x \in [0, a]$, and therefore h is a one one function; it also shows that when $x = 0$, then $u = 0$, and when $x = a$, then $u = \dfrac{\pi}{2}$; that is,

$$h(0) = 0$$

$$h(a) = \frac{\pi}{2}.$$

Substituting this information about h and k into Equation (2), we obtain:

$$\int_0^a x \longmapsto b\sqrt{1 - \frac{x^2}{a^2}} = \int_0^{\pi/2} (u \longmapsto b \cos u) \times (a \cos u)$$

since $\sqrt{1 - \dfrac{x^2}{a^2}} = \cos u$, and $Dk(u) = a \cos u$.

This integral simplifies to:

$$ab \int_0^{\pi/2} u \longmapsto \cos^2 u.$$

This is still not a standard integral, but at least we have got rid of the square root.

To complete the evaluation of the integral, we should like to use the standard forms for integrating sines and cosines. This is not immediately possible because the cosine is squared, so the first step is to express

$\cos^2 u$ in terms of a cosine that is not squared. It is one of the very useful properties of the trigonometric functions that this is possible. We use the identities*

$$\left.\begin{array}{r} 1 = \cos^2 u + \sin^2 u \\ \cos 2u = \cos^2 u - \sin^2 u \end{array}\right\} \quad (u \in R).$$

Adding and dividing by 2 gives

$$\cos^2 u = \tfrac{1}{2} + \tfrac{1}{2}\cos 2u \qquad\qquad (u \in R) \qquad\qquad \text{Equation (3)}$$
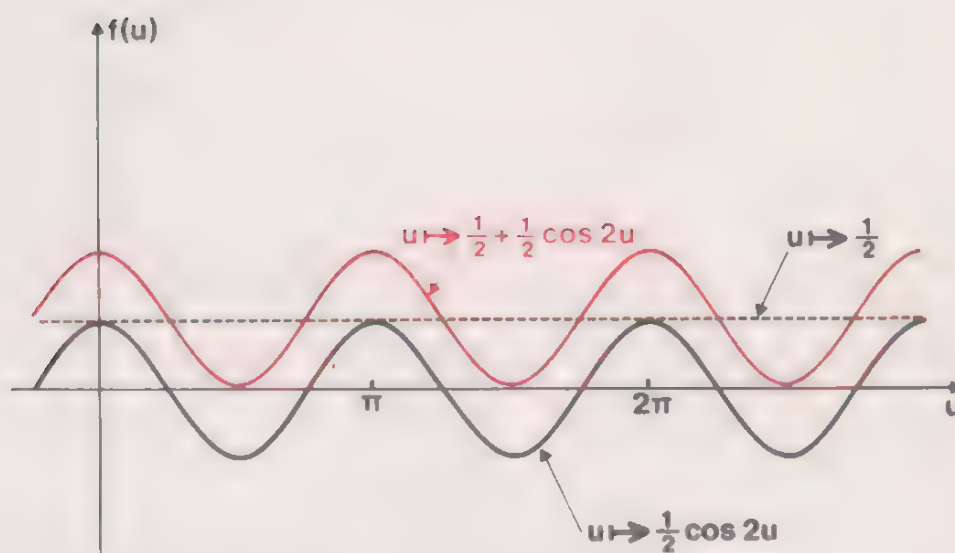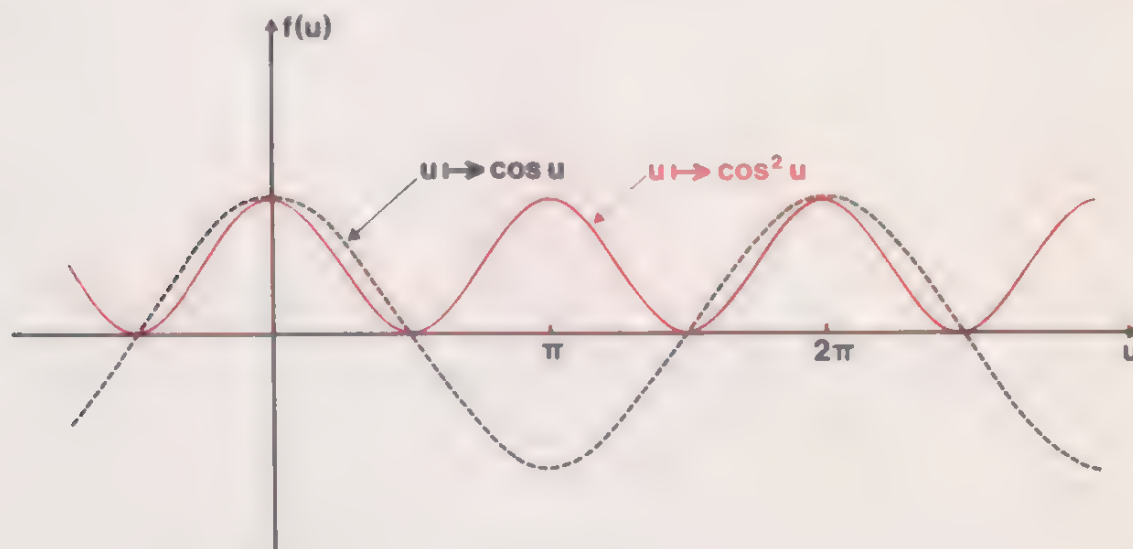
Exercise 1

Complete the evaluation of the integral, and hence of the area of the ellipse. Verify your formula by considering the special case when $a = b$.

* An *identity* (in this context) is a formula, such as $f(x) = g(x)$, that connects images under two functions and holds for all elements in their common domain (as opposed to an equation, which holds for only a few special values).

4.7   Additional Exercises
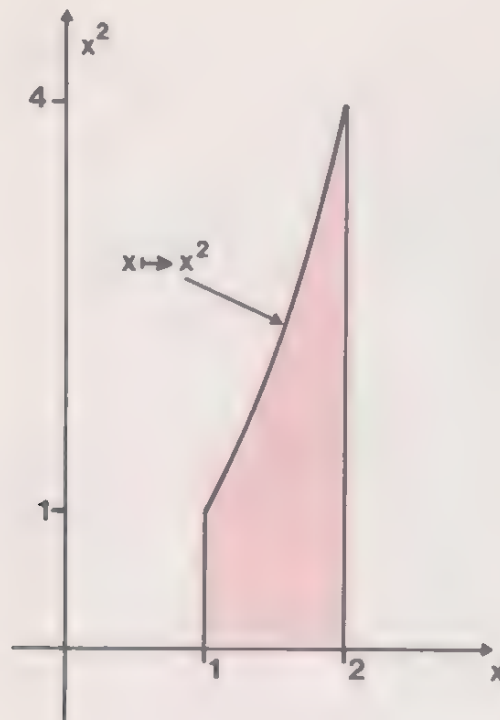
Exercise 1

Find the volume within the surface formed by rotating the graph of

$$x \longmapsto x^2 \qquad (x \in [1, 2])$$

about the x-axis.

Exercise 2

Verify that

$$\int_0^1 x \longmapsto \frac{1}{1 + x^2} = \frac{\pi}{4}.$$

using the substitution

$$x = \tan u \qquad (x \in [0, 1]),$$

with u an angle in the first quadrant. You will need to use the identity

$$1 + \tan^2 u = \sec^2 u \qquad (u \in R).$$

Exercise 3

Calculate a value for π by applying Simpson's rule with four strips to the integral in Exercise 2.

Simpson's rule for four strips is

$$\int_a^b f = \frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + y_4)$$

where h is the interval width and $y_0, y_1, \ldots, y_4$ are the ordinates at the ends of the intervals.

Work to four places of decimals.

4.8   Answers to Exercises

Section 4.1

Exercise 1
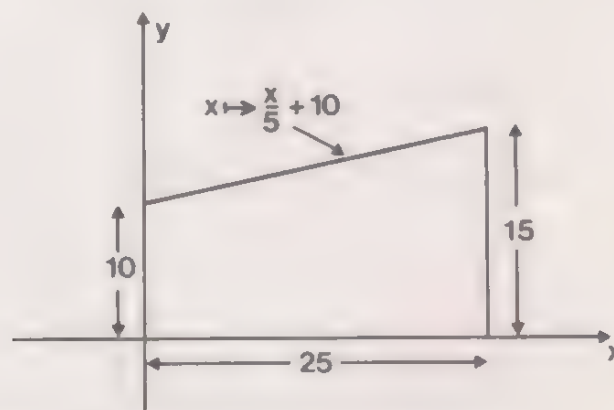
The equation of the bounding curve is given by

$$\frac{y - 10}{15 - 10} = \frac{x - 0}{25 - 0}$$

$$y = \frac{x}{5} + 10$$

Therefore the volume expressed as

$$\pi \int_0^{25} x \longmapsto \left(\frac{x}{5} + 10\right)^2$$

$$= \pi \int_0^{25} x \longmapsto \left(\frac{x^2}{25} + 4x + 100\right)$$

$$= \pi \left\{ \frac{1}{25} \frac{(25^3 - 0^3)}{3} + 4\frac{(25^2 - 0^2)}{2} + 100(25 - 0) \right\}$$

$$= \pi \left\{ \frac{19 \times 625}{3} \right\}$$

$$= 12\,440 \text{ cm}^3$$

$$= 12.44 \text{ litres (to 4 significant figures)}$$

Section 4.2

Exercise 1

$$\text{Average} = \frac{1}{4-0}\int_0^4 x \longmapsto x^2 = \frac{1}{4}\left(\frac{4^3 - 0^3}{3}\right) = 5\tfrac{1}{3}$$

Section 4.3

Exercise 1

Distances are represented by areas on the following diagram, as in the diagram for the simple example of the hiker.
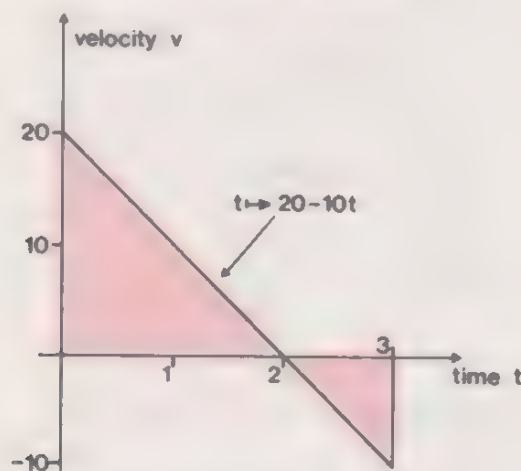
We could in this case use the formula for the area of a triangle, but we will apply integration techniques to illustrate the general method.

The graph crosses the t-axis at $t = 2$, indicating that after 2 seconds the velocity of the ball is zero. Subsequently the velocity is negative, i.e. the ball is returning to earth, so the ball will be at its highest point after 2 seconds. To calculate the distance travelled up to this time we evaluate

$$\int_0^2 t \longmapsto (20 - 10t) = 20$$

So the ball is 20 m above its starting point after 2 seconds.

In the next second it travels a distance given by

$$\int_2^3 t \longmapsto (20 - 10t) = -5$$

the negative sign indicating, as expected, that the ball is returning towards the ground during this period.

The answers are, therefore:

(i) $(20 - 5)\,\text{m} = 15\,\text{m}$,

(ii) $(20 + 5)\,\text{m} = 25\,\text{m}$.

The answer to (i), 15 m, is the result of the evaluation of

$$\int_0^3 (t \longmapsto (20 - 10t))$$

that is, the definite integral represents the distance from the starting point. Notice the more general point here that, if we form the definite integral over a domain for which the image changes sign, we need not split the domain up into sub-domains in which the numerical value is wholly positive and wholly negative. We need only do this in cases where the physical requirements of the problem demand it.

Section 4.4

Exercise 1

Your answer should contain the following:

e^{-x^2} decreases as x increases in $[0, 2]$

$$n = \frac{2}{h} \geqslant \frac{1 - e^{-4}}{5 \times 10^{-4}} \simeq 2000 \text{ intervals}$$

If the error in each ordinate is ε, then the error in the result is 2ε.

So we could use 2500 intervals (producing a possible error of 4×10^{-4}) with data accurate to 4 decimal places (producing a possible error of 1×10^{-4}).

Section 4.6

Exercise 1

Substituting from Equation (3), page 107, into the integral, and then using the rules for sums and constant factors, we obtain:

$$\tfrac{1}{2}ab \int_0^{\pi/2} u \longmapsto 1 + \tfrac{1}{2}ab \int_0^{\pi/2} u \longmapsto \cos 2u.$$

The first integral is now a standard integral and the second is almost in standard form; the simplest way to evaluate it is to note that

$$D(u \longmapsto \sin 2u) = u \longmapsto 2\cos 2u$$

so that $D(u \longmapsto \frac{1}{2}\sin 2u) = u \longmapsto \cos 2u$ and hence $u \longmapsto \frac{1}{2}\sin 2u$ is a suitable primitive.

The integral thus becomes

$$\tfrac{1}{2}ab[u \longmapsto u]_0^{\pi \cdot 2} + \tfrac{1}{2}ab[u \longmapsto \tfrac{1}{2}\sin 2u]_0^{\pi \cdot 2} = \tfrac{1}{4}\pi ab + 0.$$

This is the area of the quarter-ellipse; so we conclude that

area of ellipse $= \pi ab$

(We are not suggesting that this is the best way to calculate the area of an ellipse; if you have a feeling for geometry you may like to try to think of a more obvious method. Our main purpose here is to illustrate the method of integration by substitution.)

The check is to consider the special case where $b = a$, in which case the ellipse is a circle of radius a, whose area is correctly given by the formula above as πa^2.

Section 4.7

Exercise 1

$$\text{Volume} = \pi \int_1^2 x \longmapsto (x^2)^2$$

$$= \frac{\pi(2^5 - 1^5)}{5}$$

$$= \frac{\pi 31}{5}$$

Exercise 2

The rule of integration by substitution tells us that

$$\int_0^1 x \longmapsto \frac{1}{1 + x^2} = \int_{h(0)}^{h(1)} \left(u \longmapsto \frac{1}{1 + k(u)^2} \right) \times Dk(u).$$

Taking $x = \tan u$, we have:

$$k : u \longmapsto \tan u \qquad \left(u \in \left[0, \frac{\pi}{4} \right] \right).$$

$$h : x \longmapsto \text{the angle in } \left[0, \frac{\pi}{4}\right] \text{ whose tangent is } x$$

$$(x \in R \text{ and } x \geqslant 0),$$

and, in particular, since we have restricted u to the first quadrant:

$$h(0) = 0$$

$$h(1) = \frac{\pi}{4}, \quad \text{since} \quad \tan \frac{\pi}{4} = 1.$$
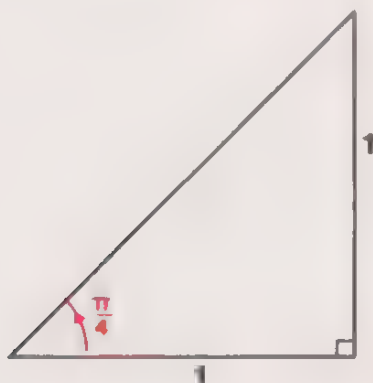
The integration therefore gives:

$$\int_0^1 x \longmapsto \frac{1}{1 + x^2} = \int_0^{\pi/4} \left(u \longmapsto \frac{1}{1 + \tan^2 u} \right) \times \sec^2 u$$

$$= \int_0^{\pi/4} u \longmapsto 1$$

$$= [u \longmapsto u]_0^{\pi/4}$$

$$= \frac{\pi}{4}.$$

Exercise 3

x	$1 + x^2$	$\dfrac{1}{1 + x^2}$
0	1.0000	1.0000
0.25	1.0625	0.9412
0.5	1.2500	0.8000
0.75	1.5625	0.6400
1.00	2.0000	0.5000

By Simpson's rule, we find (since the interval width is 0.25)

$$\frac{\pi}{4} = \int_0^1 x \longmapsto \frac{1}{1+x^2}$$

$$\simeq \frac{0.25}{3}(1.0000 + 4 \times 0.9412 + 2 \times 0.8000$$

$$+ 4 \times 0.6400 + 0.5000)$$

$$= \frac{0.25}{3}(9.4248);$$

so

$$\pi \simeq \tfrac{1}{3}(9.4248) = 3.1416,$$

which is a surprisingly accurate result from so little work!

CHAPTER 5 TAYLOR APPROXIMATIONS

5.0 Introduction

In this chapter we consider the problem of evaluating the images of real functions which cannot be expressed in terms of the elementary operation of arithmetic. The essence of the method described is to replace the function under consideration by a polynomial function in such a way that the images under the latter are a "good approximation" to the images under the former.

There are a number of ways of finding such polynomial approximations; the one we describe is particularly useful in that it only requires a knowledge of the image of the original function at one point, together with its derivative, and perhaps also some higher derivatives at the same point.

The general method is called the *Taylor approximation* or *Taylor expansion*. We shall develop the general result by way of some particular cases discussed in the early section.

The Taylor expansion has other uses than the computational use described above, and these other uses (some of which we described in this chapter) are probably more important. The reason we have chosen to introduce the subject by way of numerical approximations is that this is the simplest motivation.

5.1 The Tangent Approximation

The simplest form of Taylor's approximation to a given real function f is by a linear polynomial function of the form

$$x \longmapsto b_0 + b_1 x$$

That is, our approximation has the form

$$f(x) \simeq b_0 + b_1 x$$

Geometrically we are replacing the graph representing $y = f(x)$ by the straight line graph representing $y = b_0 + b_1 x$. Obviously there are many ways of choosing this straight line, or what is the same thing, choosing b_0 and b_1, but the Taylor approximation determines a unique straight line as follows.

Consider a straight line chosen to pass through two points on the graph of the function, one of which has coordinates $(0, f(0))$. (In the diagram $f(0) = 0$).
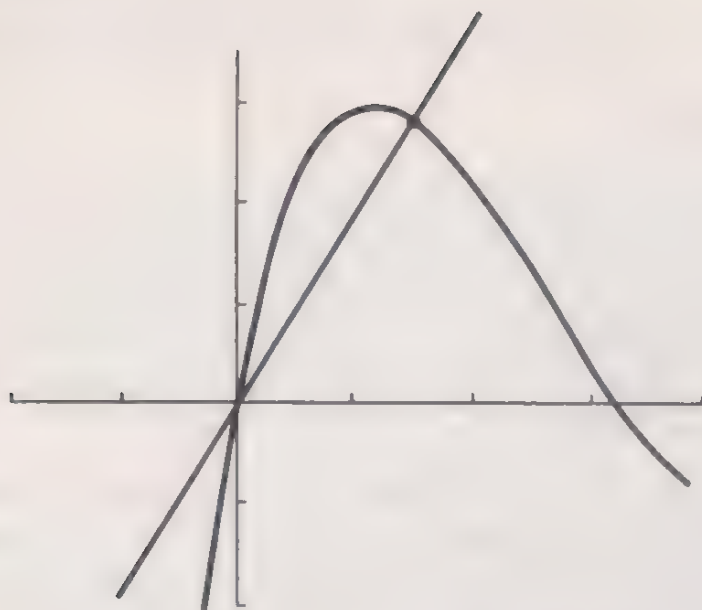
In the limit as the second point approaches $(0, f(0))$ along the curve, the straight line approaches the tangent (we are of course assuming that the limiting process is valid and we shall go on making this assumption throughout this chapter).
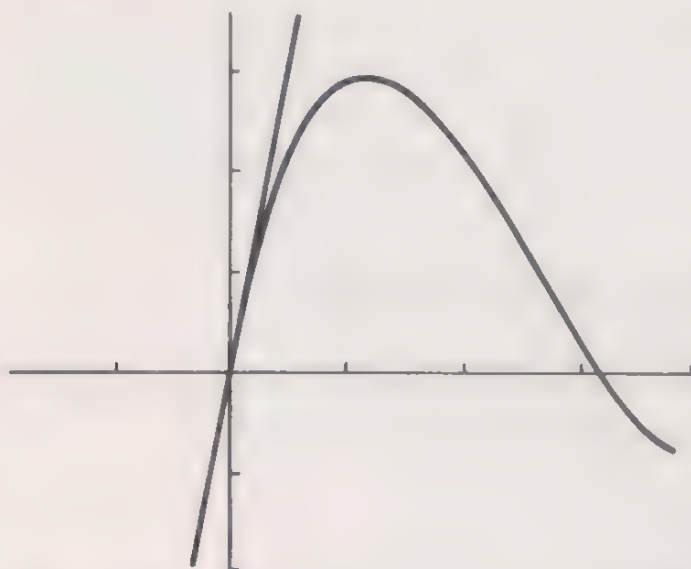
This tangent is the unique line which we require for the Taylor approximation. That is, it is the line we are looking for with equation $y = b_0 + b_1 x$. Because in the special case just described the tangent passes through $(0, f(0))$ the approximation

$$f(x) \simeq b_0 + b_1 x$$

is called the tangent approximation about $x = 0$. It is suitable for estimating $f(x)$ close to $x = 0$.

As an illustration we now calculate an approximation to $\sin\left(\dfrac{\pi}{10}\right)$.

We find the equation of the tangent to the sine curve at the origin. Since sin (0) = 0 our tangent passes through the origin.

Hence, if we assume that the equation of the tangent is $y = b_0 + b_1 x$, then we find $b_0 = 0$. Further, the derived function of sin is cos, and cos 0 = 1, so that the slope of the tangent, b_1, is 1. Hence the equation of the tangent at the origin is

$$y = x$$

and our first (tangent) approximation to sin x is

$$\sin x \simeq x.$$

Thus

$$\sin\left(\frac{\pi}{10}\right) \simeq \frac{\pi}{10} = 0.3142.$$

(The correct value, to four decimal places, is 0.3090.)
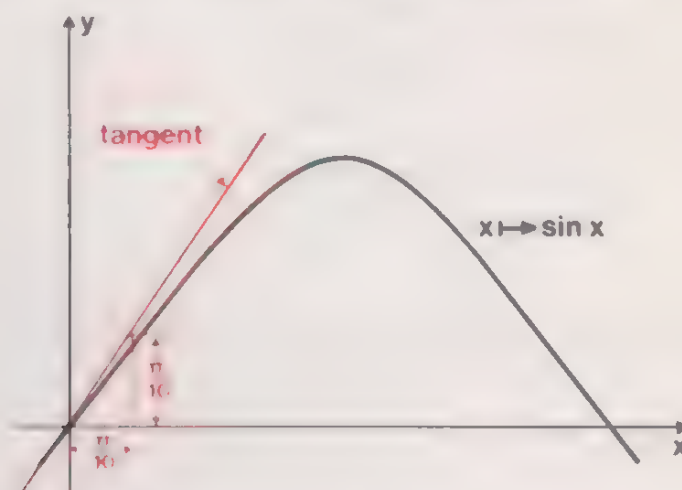
The same method can be applied to any real function f at any point $(a, f(a))$.
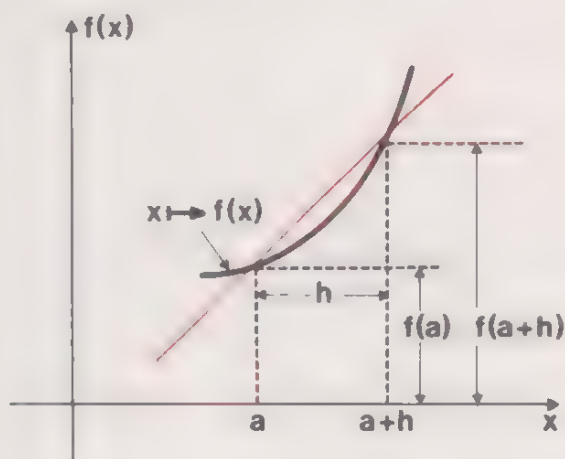
The slope of the line joining the two points is $\dfrac{f(a+h)-f(a)}{h}$, and since the line passes through $(a, f(a))$ its equation is

$$y = f(a) + \frac{f(a+h)-f(a)}{h}(x-a).$$

We want the limiting form of this equation as the right-hand point approaches the left-hand one, that is, as h approaches zero. In this limit, the line still passes through the point $(a, f(a))$, but its slope is now the derivative at this point, namely $f'(a)$.
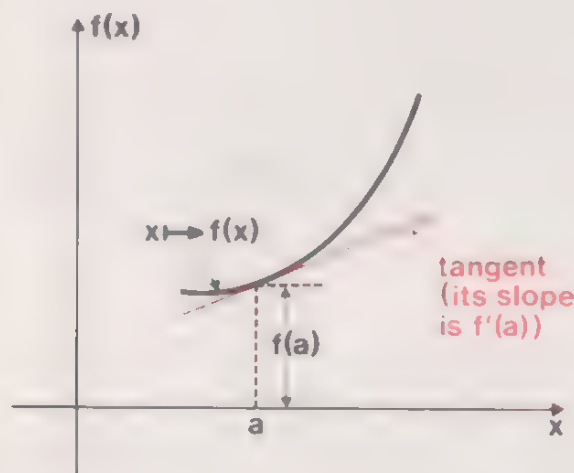
Accordingly, the equation of the tangent at $(a, f(a))$ is

$$y = f(a) + f'(a)(x-a). \qquad \text{Equation (1)}$$

The tangent approximation about $x = a$ is obtained by taking the right-hand side of Equation (1) as an approximation to $f(x)$; that is,

$$f(x) \simeq f(a) + f'(a)(x-a).$$

Exercise 1

Find the equation of the tangent to the sine curve at the point $\left(\dfrac{\pi}{6}, \sin\dfrac{\pi}{6}\right)$, and use this as an approximation to estimate $\sin\left(\dfrac{\pi}{10}\right)$. (You may assume that $\sin\left(\dfrac{\pi}{6}\right) = 0.5000$ and $\cos\left(\dfrac{\pi}{6}\right) = 0.8660$.)

Exercise 2

When a solid is heated it expands. The volume coefficient of thermal expansion of a solid may be defined as

$$\frac{\text{increase in volume due to a temperature increase of one degree}}{\text{original volume}}$$

and the linear coefficient of expansion may be defined as

$$\frac{\text{increase in any linear dimension due to a temperature increase of one degree}}{\text{original linear dimension}}.$$

For copper, the volume coefficient of expansion is about 50×10^{-6} per degree Centigrade, and the linear coefficient is 16×10^{-6} per degree Centigrade, which is about one third of the volume coefficient. Is this simple relation between the coefficients merely a coincidence?

5.2 Convergence of an Iterative Method

Although the tangent approximation is not very accurate, it is simple to use, and can be very effective when the accuracy it offers is sufficient for the matter in hand. Before proceeding to discuss how we can improve its accuracy, we shall consider how it can be used when solving equations.

An iterative method for solving equations of the form

$$x = F(x) \qquad \text{Equation (1)}$$

involves constructing a sequence $u_1, u_2, u_3, \ldots$ in which the first term is any crude approximation to a solution of Equation (1), and the later terms are calculated using the recurrence formula:

$$u_k = F(u_{k-1}) \qquad (k = 2, 3, 4, \ldots).$$

It can be shown that, if this sequence converges to a limit a, and if F is continuous at a, then a is a solution of Equation (1). To avoid wasting time calculating the elements of non-convergent sequences, it is useful to have a simple criterion by which we can tell, without actually doing the calculation, which solutions of Equation (1) (if any) can be found by this method.

There is a simple criterion, making use of the tangent approximation. To start with, we suppose that the sequence $u_1, u_2, \ldots$ does converge, and

that its limit is a. Then, for large k, the numbers u_k are close to a, and so it is natural to consider using the tangent approximation to simplify the right-hand side of the recurrence formula

$$u_k = F(u_{k-1}) \qquad (k = 2, 3, \ldots).$$

The tangent approximation for $F(u_{k-1})$ that is useful when u_{k-1} is close to a is

$$F(u_{k-1}) \simeq F(a) + F'(a)(u_{k-1} - a)$$

(see Equation 5.1.1).

Substituting this into Equation (2), we obtain

$$u_k \simeq F(a) + F'(a)(u_{k-1} - a).$$

Since a is a solution of the equation $x = F(x)$, we have $a = F(a)$, and so this last approximation is equivalent to

$$u_k - a \simeq F'(a)(u_{k-1} - a).$$

That is, when k is large, the deviation of the kth term, u_k, from the limit, a, is $u_k - a$, and differs from the preceding deviation, $u_{k-1} - a$, by a factor $F'(a)$, which is independent of k; that is,

$$k\text{th deviation} \simeq F'(a) \times ((k-1)\text{th deviation}).$$

It follows that, when k is large, the deviations will increase as k increases if $|F'(a)| > 1$. But if the sequence $u_1, u_2, \ldots$ converges to a limit a, then the deviations from a must eventually decrease as we take elements later and later in the sequence. Thus if the iterative sequence converges to a, then $|F'(a)| < 1$.

Notice the *if* in that last sentence. To complete the criterion it would be nice to be able to prove the converse statement: "if $a = F(a)$ and $|F'(a)| < 1$, then the iterative sequence converges to a". It is not quite as simple as this, however; for example, there might be two different numbers a_1 and a_2, both being solutions of $x = F(x)$ and such that $|F'(a_1)| < 1$ and $|F'(a_2)| < 1$, but the sequence could not possibly converge to both of them, since the limit of a convergent sequence is *unique*. What we can say is that if $a = F(a)$ and $|F'(a)| < 1$, and u_1 is chosen close enough to a, then the sequence $u_1, u_2, \ldots$ will converge to a, for then the deviations $u_1 - a, u_2 - a, \ldots$ approximately form a geometric progression converging to zero. If, however, u_1 is chosen so far from a that the tangent approximation for $F(u_1)$ is very inaccurate, then we have no reason to expect the sequence to converge to a. It may ultimately converge to a

anyway, but it may converge to some other solution of $x = F(x)$, or it may not converge at all.

Exercise 1

The equation $x = x^2 + \frac{1}{2}x$ has two solutions. Without calculating iterative sequences, predict which of them can be computed using the iterative method based on the recurrence formula

$$u_k = u_{k-1}^2 + \frac{1}{2}u_{k-1}.$$

5.3   The Newton–Raphson Process

As our second application of the tangent approximation in numerical methods, we shall use it to obtain a method for the numerical solution of equations which has very good convergence properties. The new method, known as the Newton–Raphson process, is again an iterative method, but here the tangent approximation is an integral part of the method, instead of being used almost as an afterthought to discuss the convergence.

Since we are not now interested in the iteration $u_n = F(u_{n-1})$, we shall not write the equation to be solved in the form $x = F(x)$, but in the more convenient form $f(x) = 0$. (The previous equation, $x = F(x)$, can be put into this form by taking $f(x) = x - F(x)$.) To construct the recurrence formula for the Newton–Raphson iteration, suppose that, after $k - 1$ steps of the iteration, the latest approximation to the solution of $f(x) = 0$ is u_k; we use the tangent approximation to f near u_k to estimate the value of x where $f(x) = 0$, and we take this estimate as our next approximation, u_{k+1}. The calculations are illustrated in the figure:
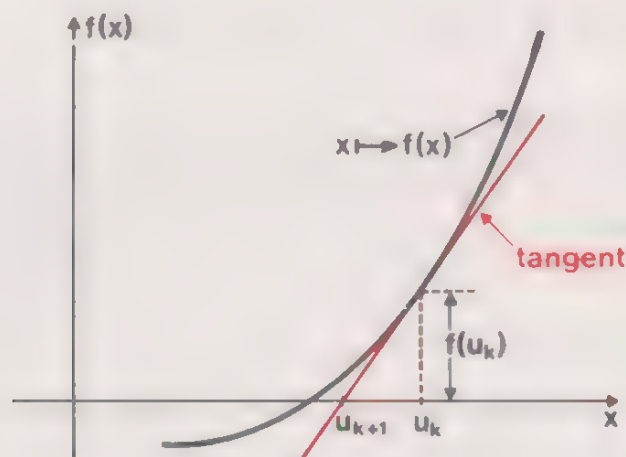
By the tangent approximation formula, Equation 5.1.1, the tangent at $(u_k, f(u_k))$ has the equation:

$$y = f(u_k) + f'(u_k)(x - u_k).$$ Equation (1)

While it may not be possible to solve the equation $f(x) = 0$ exactly (that is why we need numerical methods at all), there is no difficulty in solving the equation

(linear approximation to $f(x)$) $= 0$, Equation (2)

because it is *linear*. Using the linear approximation on the right-hand side of Equation (1) in Equation (2), we obtain

$$f(u_k) + f'(u_k)(x - u_k) = 0$$

and the solution for x is

$$x = u_k - \frac{f(u_k)}{f'(u_k)}.$$

This is the value of x where the tangent approximation to $f(x)$ is 0, and so we use it as our next approximation to the value of x where $f(x)$ itself is 0. Accordingly, the recurrence formula for the Newton Raphson method is

$$u_{k+1} = u_k - \frac{f(u_k)}{f'(u_k)}.$$

Exercise 1

Write down the Newton–Raphson recurrence formula for the equation $x^2 - a = 0$.

5.4   The Quadratic Taylor Approximation

The iterative methods for solving $f(x) = 0$ based on the tangent approximation can usually yield any degree of accuracy provided we iterate for long enough. But they only work if we can calculate $f(x)$ for any x: they do not give us a method for calculating $f(x)$ itself. We have not yet found a way of calculating $\sin\left(\dfrac{\pi}{10}\right)$ to better than about 3% accuracy.

The tangent approximation is very good near the point where the tangent touches the curve, but the accuracy falls off rapidly away from this point;

$\dfrac{\pi}{10}$ is too far away from the points of contact in the tangent approxima-

tions which we have tried for $\sin\left(\dfrac{\pi}{10}\right)$. To improve the accuracy we need

something better than the tangent approximation. One way to try to improve the approximation is to use quadratic, or even higher-degree polynomials, in place of the linear one we have been using so far.

We can obtain the quadratic approximation by fitting a quadratic function of the form

$$x \longmapsto c_0 + c_1 x + c_2 x^2,$$

where c_0, c_1, c_2 are numbers, to the given function at equally spaced points in the domain, and then making the spacing h between these points extremely small. In the limit as h approaches 0, this quadratic approximation becomes the *quadratic Taylor approximation*.

The graph of this limiting quadratic function will touch the graph of the original function; it seems evident, and could be proved, that the graph of this quadratic function not only has the same slope (first deriva- tive) but also has the same second derivative as the original function at the point of contact. Denoting the quadratic function which we are using to approximate to f by q, and the value of x where the curves touch by a, the conditions to be satisfied are

equal images for a: $\qquad\qquad\qquad q(a) = f(a)$

equal slopes at a: $\qquad\qquad\qquad q'(a) = f'(a)$ $\qquad$ **Equations (1)**

equal second derivatives at a: $\quad q''(a) = f''(a)$

These three conditions provide just sufficient information to determine the three coefficients c_0, c_1, c_2 in the expression $q(x) = c_0 + c_1 x + c_2 x^2$. The neatest way to use these conditions is to write the quadratic in the alternative form (analogous to the formula $f(a) + f'(a)(x - a)$ for the tangent approximation):

$$q(x) = b_0 + b_1(x - a) + b_2(x - a)^2. \qquad\qquad \textbf{Equation (2)}$$

Differentiating, we get

$$q'(x) = b_1 + 2b_2(x - a)$$

$$q''(x) = 2b_2$$

so that the values of the quadratic function and its derivatives at a are

$$q(a) = b_0$$

$$q'(a) = b_1$$

$$q''(a) = 2b_2.$$

Comparing with Equations (1) we find that

$$b_0 = f(a)$$

$$b_1 = f'(a)$$

$$b_2 = \tfrac{1}{2}f''(a)$$

so that, substituting in Equation (2), the quadratic Taylor approximation to $f(x)$ (when x is close to a) is

$$f(x) \simeq f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2.$$

It is usually called the quadratic Taylor approximation to f *about* a. With this formula we can take any real function f and any element a in the domain of f, for which $f(a)$ and its first two derivatives are known, and calculate an approximation to the image value of any other element x close to a in the domain.

Exercise 1

Use the quadratic Taylor approximation with $a = 0$ to estimate $\sin\left(\dfrac{\pi}{10}\right)$. In Exercise 5.1.1, the tangent approximation at $\dfrac{\pi}{6}$ gave us a 3% error in $\sin\left(\dfrac{\pi}{10}\right)$. What is the error in this new approximation? (The true value of $\sin\left(\dfrac{\pi}{10}\right)$ is 0.3090 to 4 decimal places.)

5.5 The General Taylor Approximation

In the previous section we showed how the quadratic Taylor approximation gives, in general, a better approximation than the linear one (the tangent approximation); but for some purposes even the quadratic approximation is not adequate. To look for even better approximations, it is natural to try the same method with a cubic polynomial, or one of even higher degree.

In this section we formulate the Taylor approximation that uses a polynomial of any degree, say the nth. By analogy with the method that worked for the quadratic polynomial, let us write the polynomial of degree n in the form:

$$p(x) = b_0 + b_1(x - a) + b_2(x - a)^2 + \cdots + b_n(x - a)^n$$

where $b_0, b_1, \ldots, b_n$ are numbers. (At this stage there is no reason to assume any connection between the numbers b_0, b_1 and b_2 in this section and the coefficients of the quadratic polynomial in the preceding section, but we shall see presently that they are in fact the same.) How do we determine the numbers $b_0, \ldots, b_n$? Since there are $n + 1$ of them, we need $n + 1$ conditions to fix them all. From the previous section we already have three conditions

$$p(a) = f(a)$$

$$p'(a) = f'(a)$$

$$p''(a) = f''(a)$$

where f is the function we are trying to approximate. It is natural to impose the remaining conditions by continuing the list:

$$p'''(a) = f'''(a)$$

$$p^{(4)}(a) = f^{(4)}(a)$$

$$\cdots$$

$$p^{(n)}(a) = f^{(n)}(a)$$

where $f^{(n)}(a)$ means the nth derivative of f at a. The complete list gives us exactly $n + 1$ conditions, and it is plausible to use these to determine the numbers $b_0, \ldots, b_n$ in the definition of p. The next exercise deals with the determination of these numbers.

Exercise 1

If c is a polynomial function defined by

$$c(x) = b_0 + b_1(x - a) + b_2(x - a)^2 + b_3(x - a)^3$$

and $c(a)$ and the first three derivatives $c'(a)$, $c''(a)$ and $c'''(a)$ are equal to $f(a)$, $f'(a)$, $f''(a)$ and $f'''(a)$ respectively, find b_0, b_1, b_2 and b_3, and hence

write down a formula giving c in terms of f and its first three derivatives at a.

The formula for the nth degree Taylor polynomial approximation can be calculated by writing it in the form

$$p(x) = b_0 + b_1(x - a) + b_2(x - a)^2 + \cdots + b_n(x - a)^n$$

and using the $n + 1$ conditions

$$p(a) = f(a), p'(a) = f'(a), \ldots, p^{(n)}(a) = f^{(n)}(a)$$

to determine the $n + 1$ numbers $b_0, b_1, \ldots, b_n$.

We thus obtain the Taylor approximation of degree n:

$$f(x) \simeq f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2 +$$

$$\cdots + \frac{1}{k!}f^{(k)}(a)(x - a)^k + \cdots + \frac{1}{n!}f^{(n)}(x - a)^n$$

general term

This is usually referred to as Taylor's approximation to f *about* a. The factorials* in the denominator arise because the kth derived function of $x \longmapsto (x - a)^k$ is $x \longmapsto k!$.

The value of a for which this approximation is simplest is usually 0, and the resultant form of the Taylor approximation is common enough to have a special name: it is called the Maclaurin approximation. Its formula is

$$f(x) \simeq f(0) + f'(0)x + \tfrac{1}{2}f''(0)x^2 +$$

$$\cdots + \frac{1}{k!}f^{(k)}(0)x^k + \cdots + \frac{1}{n!}f^{(n)}(0)x^n.$$

The Maclaurin approximation is a Taylor approximation to f about 0.

As an example we find the Maclaurin approximation for the case where f is the sine function. For this function the derivatives at 0 are:

$$f(0) = \quad \sin 0 = \quad 0$$
$$f'(0) = \quad \cos 0 = \quad 1$$
$$f''(0) = \quad -\sin 0 = \quad 0$$
$$f'''(0) = \quad -\cos 0 = \quad -1$$

---

* The symbol $k!$ denotes the product $1 \times 2 \times 3 \times \cdots \times k$, and is read "factorial k".

$$f^{(4)}(0) = \quad \sin 0 = \quad 0$$

$$f^{(5)}(0) = \quad \cos 0 = \quad 1$$

. . .

and the pattern $0, 1, 0, -1, 0, 1, 0, -1, \ldots$ goes on repeating itself; substituting these values into the Maclaurin approximation we get the successive approximations:

$$(n = 1 \text{ or } 2) \quad \sin x \simeq x$$

$$(n = 3 \text{ or } 4) \quad \sin x \simeq x - \frac{x^3}{3!}$$

$$(n = 5 \text{ or } 6) \quad \sin x \simeq x - \frac{x^3}{3!} + \frac{x^5}{5!}$$

$$(n = 7 \text{ or } 8) \quad \sin x \simeq x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}$$

. . . . . .

and so on.

We can compare the graphs of the polynomial approximations with the graph of the sine function. Here are some of the results:
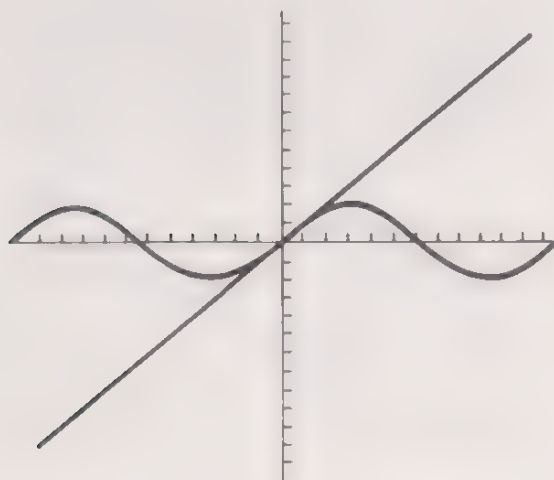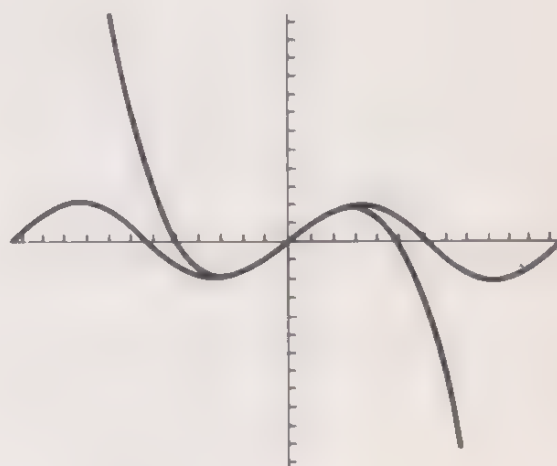
$n = 1$ $\qquad\qquad\qquad\qquad\qquad\qquad n = 3$
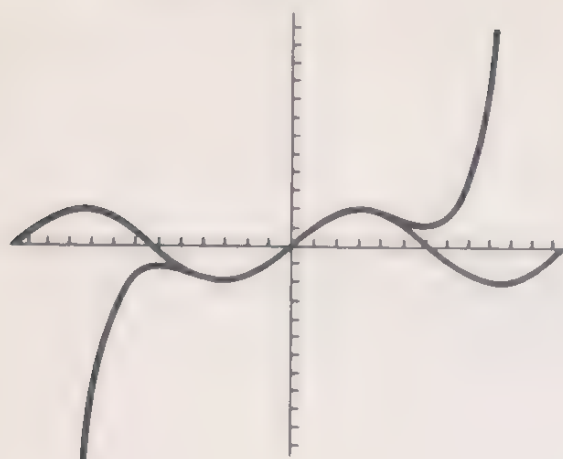
$n = 5$
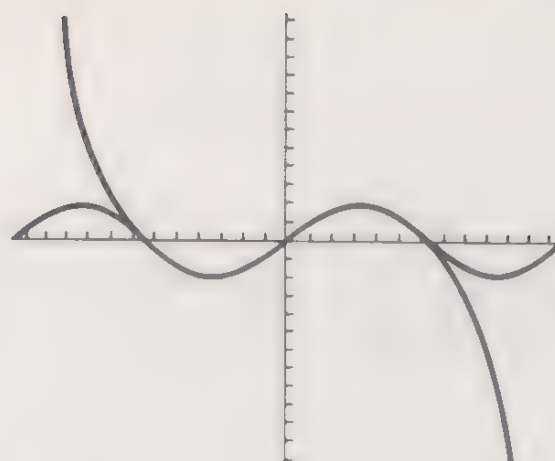
$n = 7$
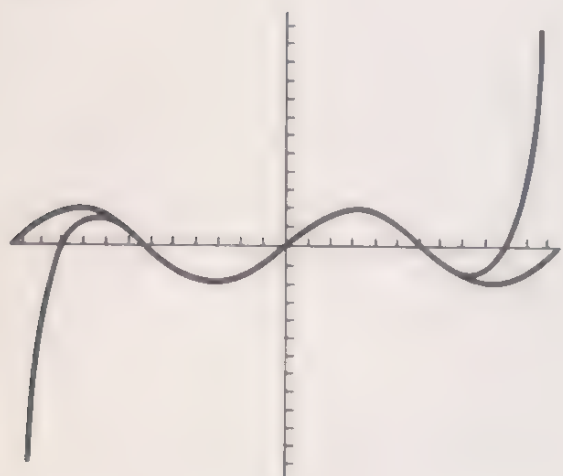
$n = 9$
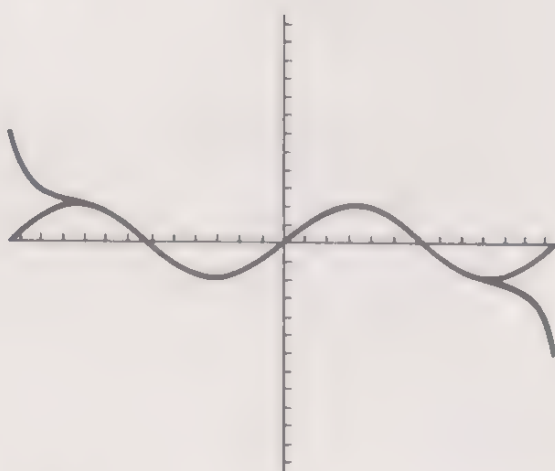
$n - 11$

As we make n larger and larger the approximation gets better and better, in the sense that the polynomial fits the sine curve over a wider and wider interval, and there seems to be no restriction on the width of the interval if we take a polynomial of sufficiently high degree. But note that the diagrams are somewhat deceptive, since the lines must have some thickness in order to be visible, and all they really show is that the approximating curve gradually "follows" the sine curve as the degree of the approximating polynomial increases.

We can also calculate the first few approximations to $\sin\left(\dfrac{\pi}{10}\right)$ and see how well these approximate to the correct value. These are

$$n = 1 \quad \sin\left(\frac{\pi}{10}\right) \simeq \frac{\pi}{10} \qquad\qquad = 0.3141593$$

$$n = 3 \quad \sin\left(\frac{\pi}{10}\right) \simeq \frac{\pi}{10} - \frac{1}{3!}\left(\frac{\pi}{10}\right)^3 = 0.3089921$$

$$n = 5 \qquad\qquad\qquad = 0.3090176$$

$$n = 7 \qquad\qquad\qquad = 0.3090170.$$

If we continued with higher values of n the result would still be 0.3090170 to 7 significant figures.

It is a remarkable fact that, knowing the images of the sine function and its derived functions at the single element 0, Maclaurin's formula gives us a method of investigating the images of *all* the real numbers under the sine function.

Exercise 2

Find the general Maclaurin approximation to the cosine function, and calculate the first three distinct Maclaurin approximations for $\cos(0.3)$ to 3 decimal places. Compare your results with the true value, 0.9553.

In the examples considered so far, Maclaurin's approximation has been extremely successful; the following exercise shows that this is not always the case.

Exercise 3

Find the general Maclaurin approximation to the function

$$x \longmapsto (1 - x)^s \qquad (x \in R, x \neq 1),$$

where s is any real number.

Do you recognize the approximation when s is a positive integer?

Calculate the first few Maclaurin approximations to $(1 - x)^{-1}$, where

(i) $x = 0.1$ \qquad (ii) $x = 10$.

Look up the answer to Exercise 3. Observe that the method was successful for $x = 0.1$, but for $x = 10$ the "approximations" bear no relation whatever to the correct value! The Taylor (Maclaurin) approximation method is quite temperamental: sometimes it is very effective, but on other occasions the approximations it produces are wide of the mark. The method is a very powerful one, but to be able to use it without getting into trouble one needs either very sound intuition or some theorems that will specify the situations in which the method is successful. In the next section of the text we shall leave the exploratory approach we have been using and look at the theory of the Taylor approximation method from a rigorous point of view.

5.6   Errors in the Taylor Approximation

The main purpose of this section is to enable you to recognize the situations where the Taylor (or Maclaurin) approximation method works satisfactorily, so that you will know how to take advantage of the method without getting false results.

To explain the principle of the method, we consider first how to estimate the error in the simplest of the Taylor polynomial approximations, the tangent approximation. The error in any approximation is defined to be

$$\text{error} = (\text{approximation}) - (\text{exact value})$$

It is a little more convenient to work not with the error itself but with its negative, which is the correction that must be added to the approximation to cancel the error and thus yield the exact value:

$$\text{correction} = (\text{exact value}) - (\text{approximation}).$$

The error and the correction have the same magnitude (modulus), so that any bound on the magnitude of the correction is automatically an error bound too. For any given function f, let us denote the correction to the tangent approximation for $f(x)$ about some given point a by $C_1(x)$, the subscript 1 indicating that this refers to the Taylor approximation of degree one. The formula for the tangent approximation (which we found in section 5.1) is

$$f(x) \simeq f(a) + f'(a)(x - a),$$

and hence the correction is given by

$$C_1(x) = f(x) - (f(a) + f'(a)(x - a)). \qquad \text{Equation (1)}$$

Now $C_1(x)$ is the number we wish to estimate, but let us first get some idea of its size by trying some suitable approximations.
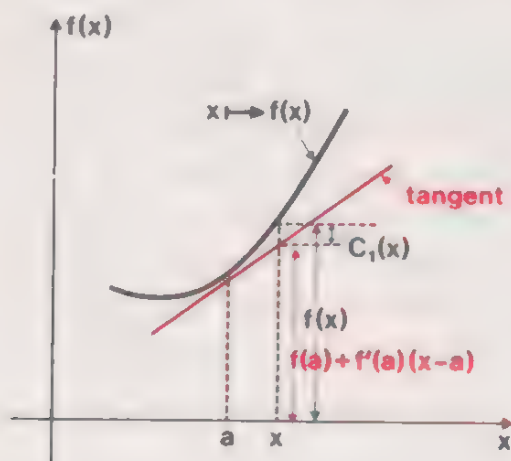
One way of getting an idea of the size of $C_1(x)$ is to replace $f(x)$ on the right of Equation (1) by a convenient approximation. What approximation would you suggest? The tangent approximation about a will not do, for that would give

$$C_1(x) \simeq (\text{tangent approx.}) - (\text{tangent approx.}) = 0$$

which is no help. But, by using the next Taylor polynomial for $f(x)$, we can get a useful estimate; it is

$$C_1(x) \simeq (f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2)$$

$$-(f(a) + f'(a)(x - a))$$

$$= \tfrac{1}{2}f''(a)(x - a)^2.$$

Thus, $C_1(x)$ is roughly proportional to the square of the distance $(x - a)$, and also to the second derivative of f at a. Both these facts can also be seen from the above figure, especially if it is redrawn to show how $C_1(x)$ depends on x.
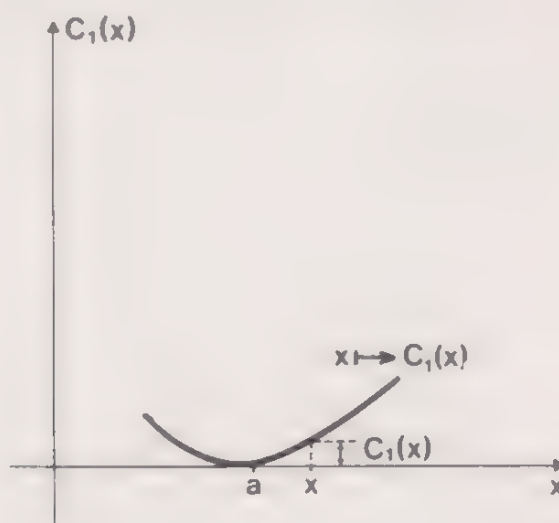
The problem now is to convert the rough estimate for the correction to the tangent approximation about a,

$$C_1(x) \simeq \tfrac{1}{2}f''(a)(x - a)^2,$$

into a precise specification of the accuracy of this approximation. The above result suggests that it may be possible to specify the accuracy of the tangent approximation by a formula such as

$$|C_1(x)| \leqslant \tfrac{1}{2}B(x - a)^2 \qquad\qquad\qquad \text{Inequality (1)}$$

in which B is somehow related to the second derived function of f.

In fact, this method of specifying the accuracy does prove satisfactory. It can be shown that the result holds provided B is an upper bound on the magnitude of the second derivative of f over the interval $[a, x]$ (or $[x, a]$ if $x < a$); that is,

provided $|f''(t)| \leqslant B$ $(t \in [a, x])$.   Inequality (2)

Inequalities (1) and (2) together constitute a statement of Taylor's Theorem

for the tangent approximation. It can be stated in the form that Inequality (2) implies Inequality (1).

Example 1

Consider $\exp x$, near $x = 1$. In this case the tangent approximation is

$$\exp x \simeq \exp(1) + (x - 1) \times \exp'(1)$$
$$= 2.7183 + (x - 1) \times 2.7183,$$

and Taylor's Theorem tells us that the correction satisfies the inequality

$$|C_1(x)| \leqslant \tfrac{1}{2} B(x - 1)^2,$$

provided B satisfies the inequality

$$|\exp t| \leqslant B \qquad (t \in [1, x])$$

(since $\exp'' = \exp$).
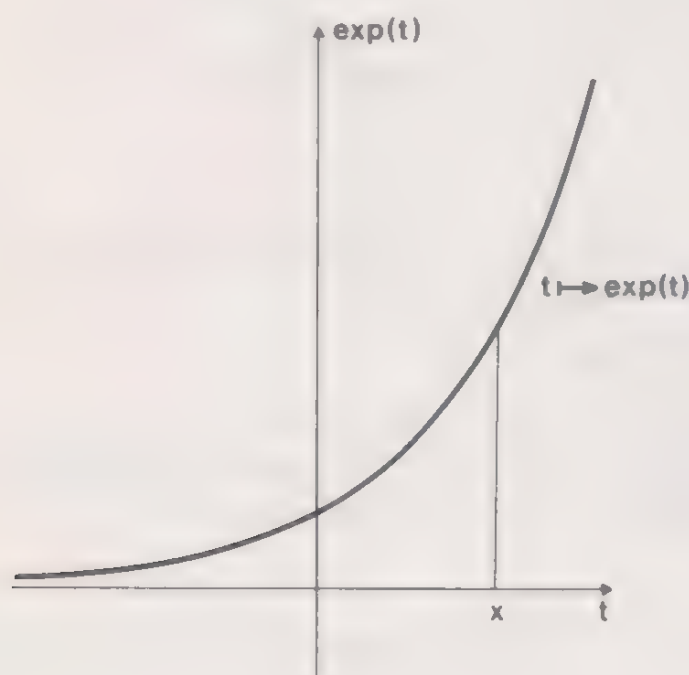
Since $\exp t$ increases as t increases, its largest value for $t \in [1, x]$ is achieved when t is the largest number in the interval $[1, x]$, which is x if $x > 1$ and 1 if $x < 1$. Accordingly we can satisfy the last inequality by taking B to be the image under the exponential function of the largest number in the interval:

$$B = \begin{cases} \exp x & \text{if } x > 1 \\ e & \text{if } x < 1. \end{cases}$$

(We could take B larger than this if we wished, and still satisfy the required inequality, but this would weaken the condition given by the first inequality without gaining anything.) Thus Taylor's Theorem tells us that

$$\exp x \simeq 2.7183 + (x - 1) \times 2.7183$$

with a correction of magnitude not exceeding

$$\begin{cases} \frac{1}{2}(\exp x) \times (x - 1)^2 & \text{if } x > 1 \\ \frac{1}{2}e \times (x - 1)^2 & \text{if } x < 1. \end{cases}$$

For example, if $x = 0.8$, Taylor's Theorem tells us that the magnitude of the correction cannot exceed

$$\frac{1}{2} \times 2.7183 \times (-0.2)^2 = 0.0544.$$

The actual correction (to the accuracy shown, as always) is

$$\exp(0.8) - (2.7183 + (-0.2) \times 2.7183)$$

$$= 2.2255 - 2.1746$$

$$= 0.0509.$$

If $x = 1.2$, Taylor's Theorem tells us that the magnitude of the correction cannot exceed

$$\frac{1}{2} \times 3.3201 \times (0.2)^2 = 0.0664.$$

The actual correction is

$$\exp(1.2) - (2.7183 + (0.2) \times 2.7183)$$

$$= 3.3201 - 3.2620$$

$$= 0.0581.$$

Thus in both cases the theorem is verified.

Exercise 1

Use Taylor's Theorem with $a = 0$ to obtain a maximum error for the approximation

$$\exp x \simeq 1 + x$$

for $x < 0$.

Deduce that

$$\exp(-0.2) \in [0.78, 0.82].$$

Exercise 2

Use Taylor's Theorem to obtain a maximum error for the tangent approximation about 0 to $\sin\left(\dfrac{\pi}{10}\right)$,

$$\sin\left(\frac{\pi}{10}\right) \simeq \frac{\pi}{10},$$

and compare it with the actual error. (The correct value of $\sin\left(\dfrac{\pi}{10}\right)$ is 0.3090 to 4 decimal places.)

5.7 The General Taylor Theorem

The Taylor approximation of degree n, obtained in section 5.5, is:

$$f(x) \simeq f(a) + f'(a)(x - a) + \cdots + \frac{1}{n!} f^{(n)}(a)(x - a)^n.$$

The correction associated with this approximation is therefore

$$C_n(x) = f(x) - \left(f(a) + f'(a)(x - a) + \cdots + \frac{1}{n!} f^{(n)}(a)(x - a)^n \right).$$

Just as in the case of the tangent approximation, we can get a rough approximation to $C_n(x)$ by using the next approximation for $f(x)$. We obtain this by replacing n by $n + 1$ in the above Taylor approximation, and we find that

$$C_n(x) \simeq \left(f(a) + f'(a)(x - a) + \cdots + \frac{1}{n!} f^{(n)}(a)(x - a)^n \right.$$

$$+ \frac{1}{(n+1)!} f^{(n+1)}(a)(x-a)^{n+1} \Bigg)$$

$$- \left(f(a) \quad f'(a)(x-a) + \cdots + \frac{1}{n!} f^{(n)}(a)(x-a)^n \right).$$

Thus $\quad C_n(x) \simeq \dfrac{1}{(n+1)!} f^{(n+1)}(a)(x-a)^{n+1}.$

This suggests that there may be a useful formula for the accuracy of the nth degree Taylor approximation of the form:

$$|C_n(x)| \leqslant \frac{1}{(n+1)!} B_{n+1} |(x-a)^{n+1}| \qquad \text{Inequality (1)}$$

where B_{n+1} depends on $f^{(n+1)}$. As in the case of the tangent approximation $(n = 1)$, it is possible to show that a sufficient condition for the above inequality to hold is

$$|f^{(n+1)}(t)| \leqslant B_{n+1} \qquad (t \in [a, x]), \qquad \text{Inequality (2)}$$

where $f^{(n+1)}$ is continuous throughout the interval $[a, x]$, and where $[a, x]$ is to be interpreted as $[x, a]$ if $x < a$.

The statement that Inequality (2) implies Inequality (1) is the general form of Taylor's Theorem.

A demonstration of Taylor's Theorem may be given as follows:

We may drop the suffix of B in inequalities (1) and (2). We then wish to show that if B satisfies the inequality

$$|f^{(n+1)}(t)| \leqslant B \qquad (t \in [a, x]), \qquad \text{Inequality (3)}$$

then $\quad |C_n(x)| \leqslant \dfrac{1}{(n+1)!} B |x-a|^{n+1} \qquad \text{Inequality (4)}$

Inequality (3) is intended to imply that the $(n+1)$th derivative of f exists at all points in $[a, x]$. We define the function

$$C_n : t \longmapsto f(t) - \left[f(a) + (t-a)f'(a) + \frac{1}{2}(t-a)^2 f''(a) + \cdots \right.$$

$$\left. + \frac{1}{n!}(t-a)^n f^{(n)}(a) \right] \qquad (t \in [a, x]).$$

This definition is consistent with the definition of $C_n(x)$ already given, and it makes sense for all points in the domain of f, since we have stipulated that f is $n + 1$ times differentiable at all points in this domain.

We shall estimate $C_n(x)$ by estimating its $(n + 1)$th derivative and then integrating $n + 1$ times. Differentiating the function C_n, we obtain:

$$C_n'(t) = f'(t) - \left[f'(a) + (t - a)f''(a) + \cdots \right.$$

$$\left. + \frac{1}{(n - 1)!}(t - a)^{n-1}f^{(n)}(a) \right]$$

$$C_n''(t) = f''(t) - \left[f''(a) + \cdots + \frac{1}{(n - 2)!}(t - a)^{n-2}f^{(n)}(a) \right]$$

$$\ldots$$

$$C_n^{(n)}(t) = f^{(n)}(t) - f^{(n)}(a)$$

$$C_n^{(n+1)}(t) = f^{(n+1)}(t)$$

where $t \in [a, x]$ in each case.

Combining the last equation with Inequality (3), we obtain the estimate

$$|C_n^{(n+1)}(t)| \leqslant B \qquad (t \in [a, x]). \qquad \text{Inequality (5)}$$

We can use this information to estimate $C_n(x)$ itself, by a succession of $n + 1$ integrations.

For simplicity we confine detailed discussion to the case where $x > a$, and to the upper bound on $C_n^{(n+1)}(t)$ implied by Inequality (5). The other cases can be treated similarly. The upper bound on $C_n^{(n+1)}(t)$ given by **Inequality (5)** is

$$C_n^{(n+1)}(t) \leqslant B \qquad (t \in [a, x]).$$

Integrating from a to s, where $s \in [a, x]$, gives (see diagram):

$$\int_a^s C_n^{(n+1)} \leqslant \int_a^s (t \longmapsto B)$$
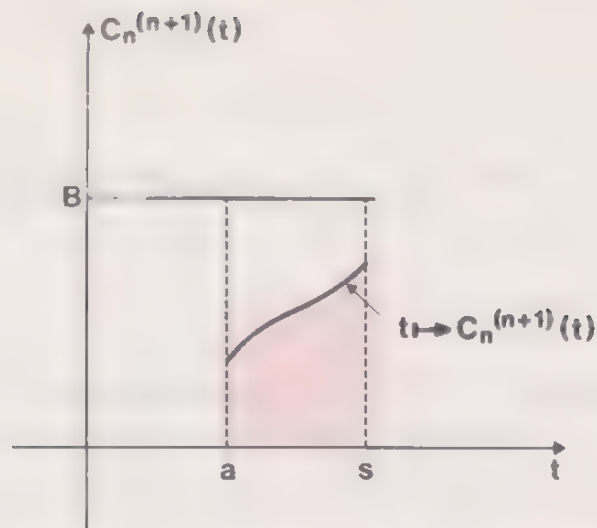
Evaluating the integrals with the help of the Fundamental Theorem of Calculus (Volume 1, Chapter 9) gives, (since $C_n^{(n+1)} = DC_n^{(n)}$, by definition)

$$C_n^{(n)}(s) - C_n^{(n)}(a) \leqslant (s - a)B.$$

But the equation which we obtained earlier for $C_n^{(n)}(t)$ shows us that $C_n^{(n)}(a) = 0$, and since the last inequality holds for all s in $[a, x]$, it follows that:

$$C_n^{(n)}(t) \leqslant (t - a)B \qquad (t \in [a, x]).$$

Now we can repeat the procedure and reduce the order of the derivative of C_n one further. Integration from a to s, with $s \in [a, x]$, gives (see diagram):

$$\int_a^s C_n^{(n)} \leqslant \int_a^s (t \longmapsto (t - a)B)$$
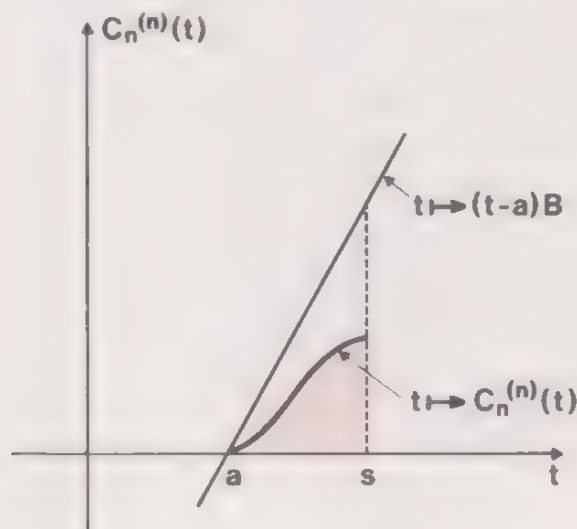
That is,

$$C_n^{(n-1)}(s) - C_n^{(n-1)}(a) \leqslant \tfrac{1}{2}(s - a)^2 B.$$

But, once again, we have $C_n^{(n-1)}(a) = 0$, and since the last inequality holds for all s in $[a, x]$, it follows that:

$$C_n^{(n-1)}(t) \leqslant \tfrac{1}{2}(t - a)^2 B \qquad (t \in [a, x]).$$

Repeating the procedure $n - 1$ more times we obtain:

$$C_n^{(n-2)}(t) \leqslant \frac{1}{3!}(t - a)^3 B \qquad\qquad (t \in [a, x]),$$

$$C_n^{(n-3)}(t) \leqslant \frac{1}{4!}(t - a)^4 B \qquad\qquad (t \in [a, x]),$$

and finally

$$C_n(t) \leqslant \frac{1}{(n + 1)!}(t - a)^{n+1} B \qquad (t \in [a, x]).$$

This is precisely the upper bound on $C_n(t)$ given by Taylor's Theorem in the case $x > a$ (since then we have $t \geqslant a$, so that $t - a$ is the same as $|t - a|$). Applying the same procedure for lower bounds, and for the case where $x < a$, we can complete the demonstration of the form of Taylor's Theorem given in the text.

Exercise 1

Write down the Maclaurin polynomial approximations of degrees 1, 2, 3 and 4 for cos x. Use Taylor's Theorem to find an integer n such that the Maclaurin polynomial of degree n for cos x gives an approximation for cos(2) that is accurate to 2 decimal places, and write down the Maclaurin polynomial of this degree. (If you find this part very difficult, Exercise 5.5.2 and its solution may be of some help.) The table of factorials may also be useful:

n	$n!$
1	1
2	2
3	6
4	24
5	120
6	720
7	5 040
8	40 320
9	362 880
10	3 628 800

5.8 Infinite Series

In this section we shall need some of the concepts from Volume 1, Chapter 6. Rather than refer back, we repeat those bits that are necessary to follow the argument.

So far we have shown how to obtain various polynomial approximations to the image of a given function, sin say, for a given element, x say, in its domain:

$$\sin x \simeq x$$

$$\sin x \simeq x - \frac{x^3}{3!}$$

$$\sin x \simeq x - \frac{x^3}{3!} + \frac{x^5}{5!} \quad \text{etc.}$$

We can show that, in favourable cases (of which this example is one) the sequence of successive approximations thus obtained converges and has the exact image value as its limit. This sequence of successive approximations is calculated by adding a term, such as $\frac{-x^3}{3!}$ or $\frac{x^5}{5!}$ to the preceding approximation. The successive terms that we may add also form an infinite sequence:

$$x, \quad \frac{-x^3}{3!}, \quad \frac{x^5}{5!}, \quad \frac{-x^7}{7!}, \ldots$$

To calculate one of the polynomial approximations to $\sin x$, we choose a positive integer n, and add up the first n consecutive members of this sequence. The more consecutive members we add in, the better is the approximation to $\sin x$. This is usually represented by writing

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

The expression on the right-hand side of this equation is called an *infinite series*. The three dots are used to indicate that the expression does not terminate.

The successive approximations:

$$x, \qquad\qquad x - \frac{x^3}{3!}, \qquad\qquad x - \frac{x^3}{3!} + \frac{x^5}{5!} \quad \text{etc.}$$

are called the *partial sums* of the infinite series. Here the sequence of partial sums converges to a limit, namely $\sin x$; this limit is called the (total) *sum* of the infinite series.

It is very important to understand just what we mean by an infinite series. We give these important definitions formally:

An infinite series is an expression of the form

$$a_1 + a_2 + a_3 + \cdots$$

The partial sums of the infinite series are the sums:

$$S_k = a_1 + a_2 + \cdots + a_k \qquad (k = 1, 2, 3, \ldots).$$

If the sequence of partial sums,

$$S_1, S_2, S_3, \ldots$$

converges to a limit S, then we say that the series converges (or is convergent) to the sum S, and we write

$$S = a_1 + a_2 + a_3 + \cdots.$$

If the sequence of partial sums does not converge, then we say that the series diverges (or is divergent): we cannot find a sum for it.

It is important to note the difference between the infinite series

$$a_1 + a_2 + a_3 + \cdots$$

and the infinite sequence

$$a_1, a_2, a_3.$$

Example 1

You may have met the formula for the sum of k terms of a geometric progression,

$$a + ar + ar^2 + \cdots + ar^{k-1} = a\left(\frac{1 - r^k}{1 - r}\right) \qquad (r \in R, r \neq 1).$$

This is the kth partial sum, S_k, of the infinite series

$$a + ar + ar^2 + \cdots.$$

This series is called the infinite geometric series; the number r is called the common ratio.

As an example, let us take $a = 1$ and $r = \frac{1}{2}$; then we have:

$$1 + \frac{1}{2} + \frac{1}{4} + \cdots + 2^{-(k-1)} = \frac{1 - 2^{-k}}{\frac{1}{2}}$$

$$= 2 - 2^{-(k-1)}.$$

Thus the sequence $S_1, S_2, S_3, \ldots$ is now

$$2 - 1, 2 - \frac{1}{2}, 2 - \frac{1}{4}, \ldots$$

which converges to 2, so we can write

$$1 + \frac{1}{2} + \frac{1}{4} + \cdots = 2.$$

Exercise 1

For what values of r can we define a sum for the infinite geometric series

$$1 + r + r^2 + r^3 + \cdots$$

and what is the formula for the sum in each case?

Exercise 2

For a given function f and given numbers x and a, if the corrections $C_k(x)$ satisfy

$$\lim_{k\,\text{large}} C_k(x) = 0,$$

what can we conclude about the infinite series:

$$f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2 + \cdots?$$

What can we conclude if the corrections do *not* satisfy the above condition?

Exercise 3

Consider the function

$$f : x \longmapsto \frac{1}{1 + x^2} \qquad (x \in R).$$
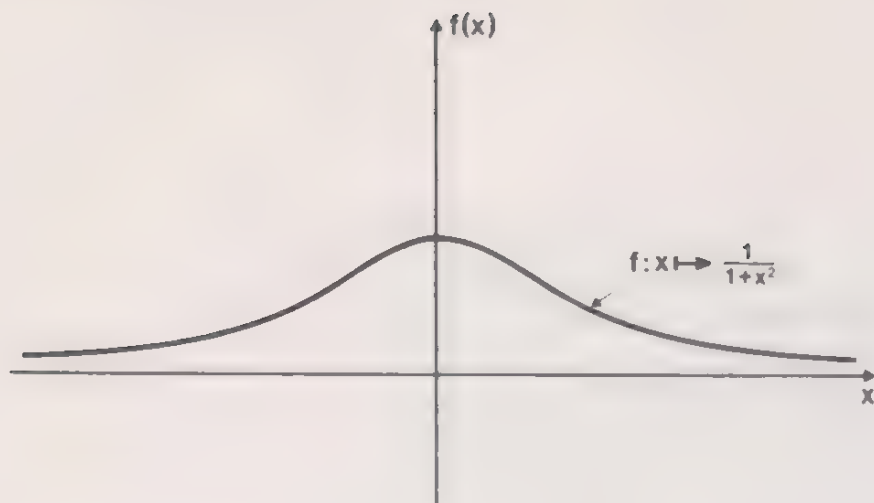
By considering the geometric series

$$1 - x^2 + x^4 - x^6 + \cdots$$

obtain a sequence of polynomial approximations for $\dfrac{1}{1 + x^2}$. (These approximations are the Maclaurin polynomials for f.) Use the results of Exercise 1 to find the set of values of x for which this sequence converges to $\dfrac{1}{1+x^2}$.

Exercise 4

In Exercise 4.7.2 we asked you to verify the formula

$$\frac{\pi}{4} = \int_0^1 x \longmapsto \frac{1}{1 + x^2}.$$

Use the approximations obtained in the preceding exercise to get a sequence of successive approximations to $\dfrac{\pi}{4}$. Assuming that this sequence really does converge to the limit $\dfrac{\pi}{4}$, write down an infinite series whose sum is $\dfrac{\pi}{4}$.

It follows from Exercise 2 that the infinite series notation provides a convenient way of summarizing the type of result obtained earlier in this chapter. For example, by writing

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots \qquad (x \in R)$$

we can concisely express a statement that would otherwise look something like this:

"the correction $C_n(x)$ to the nth degree Maclaurin approximation

$$\sin x \simeq x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^{\frac{m-1}{2}} \frac{x^m}{m!},$$

where $m = n$ if n is odd, and $m = n - 1$ if n is even, satisfies

$$\lim_{n \text{ large}} C_n(x) = 0$$

for all $x \in R$".

Similarly, by writing

$$\frac{1}{1 - x} = 1 + x + x^2 + \cdots \qquad (x \in R, |x| < 1)$$

we paraphrase the statement:

"if $|x| < 1$, then the sum $S_n(x)$ of the geometric series

$$1 + x + x^2 + \cdots + x^{n-1} \quad \text{satisfies}$$

$$\lim_{n \text{ large}} S_n(x) = \frac{1}{1 - x}".$$

To conclude this section, we summarize (for reference) a number of useful formulas of this kind.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad (x \in R);$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \qquad (x \in R);$$

$$\exp x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \qquad (x \in R);$$

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad (x \in R, |x| < 1)$$

$$(1 + x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha - 1)}{2!}x^2 + \frac{\alpha(\alpha - 1)(\alpha - 2)}{3!}x^3 + \cdots$$

$$(x \in R, |x| < 1),$$

where α is any real number. If α is a positive integer or zero, then all the terms of the last series after the $(\alpha + 1)$th are 0, so that the series reduces to a polynomial of degree α, and for these values of α the formula holds for all real x and not just for those satisfying $|x| < 1$.

5.9 Additional Exercises

Exercise 1

Apply the Newton–Raphson method to find a solution of

$$x = \sin x + \tfrac{2}{3}\pi,$$

lying between 2 and 3.

Exercise 2

Use the quadratic Taylor approximation with $a = 1$ to evaluate $\exp(1.2)$ approximately, given that $\exp(1) = e = 2.72$ to 2 decimal places.

k	$\left(1 + \dfrac{0.1}{k}\right)^k$
1	1.1
2	1.1025
3	1.1034
4	1.1038
5	1.1041
6	1.1043
7	1.1044
8	1.1045
9	1.1046
10	1.1046

Exercise 3

Find the general Maclaurin approximation to the exponential function, and calculate the first few Maclaurin approximations for $\exp(0.1)$ to 3 decimal places. Compare your results with the calculation of $\exp(0.1)$ directly from the definition of the exponential function given in Volume 1, page 104 the first 10 steps of that calculation are given in the table.

Exercise 4

Use the formula

$$C_1(x) \simeq \tfrac{1}{2}f''(a)(x - a)^2$$

to estimate (to 2 decimal places) the correction to the tangent approximation at 1 to the exponential function for $x = 0.8, 0.9, 1.1, 1.2$.

5.10   Answers to Exercises

Section 5.1

Exercise 1

From Equation (1), the equation of the tangent is

$$y = \sin\left(\frac{\pi}{6}\right) + \cos\left(\frac{\pi}{6}\right)\left(x - \frac{\pi}{6}\right)$$

since $\sin' = \cos$. The approximation to $\sin\left(\dfrac{\pi}{10}\right)$ is therefore

$$\sin\left(\frac{\pi}{10}\right) \simeq \sin\left(\frac{\pi}{6}\right) + \cos\left(\frac{\pi}{6}\right)\left(\frac{\pi}{10} - \frac{\pi}{6}\right)$$

$$= 0.5000 + 0.8660(0.3142 - 0.5236)$$

$$= 0.3187.$$

(This approximation is about 3% larger than the correct value, 0.3090.)

Exercise 2

No. Suppose we have a cube of any solid and we increase the temperature of the cube by 1 degree. If the length of the side of the cube is L, then the new length will be $L(1 + x)$, where x is the linear coefficient of expansion. Thus the new volume is $L^3(1 + x)^3$.

Thus the volume coefficient of expansion is

$$\frac{L^3(1 + x)^3 - L^3}{L^3} = (1 + x)^3 - 1.$$

Now if

$$f : x \longmapsto (1 + x)^3 - 1 \qquad (x \in R),$$

then

$$f' : x \longmapsto 3(1 + x)^2 \qquad (x \in R).$$

So the tangent approximation to $(1 + x)^3 - 1$, using the tangent at $x = 0$, is

$$(1 + x)^3 - 1 \simeq f(0) + f'(0)(x - 0)$$

$$= 3x.$$

Thus the volume coefficient of expansion is approximately three times the linear coefficient for *any* solid — and so the case of copper was not merely a coincidence.

Section 5.2

Exercise 1

Solving the quadratic equation directly, we obtain the solutions $x = 0$ and $x = \frac{1}{2}$. Convergence of the iterative sequence depends on the value of $|F'(a)| = |2a + \frac{1}{2}|$. If $a = 0$, $|F'(a)|$ is $\frac{1}{2}$, which is less than 1. So provided the initial guess is close enough to 0, the iterative method will work. If $a = \frac{1}{2}$, then $|F'(a)| > 1$, so that we cannot get the solution $x = \frac{1}{2}$ by the given recurrence formula, unless we choose a very lucky starting value. The following table gives a sample iteration.

	Sequence starting near 0	Sequence starting near $\frac{1}{2}$
u_1	0.1	0.6
u_2	0.06	0.66
u_3	0.0336	0.7656
u_4	0.0179	0.9689
u_5	0.00928	1.423
u_6	0.00473	2.738
u_7	0.00239	8.863
u_8	0.00120	82.979
	(converging to 0)	(diverging)

Section 5.3

Exercise 1

In this case $f(x) = x^2 - a$ and $f'(x) = 2x$. The Newton Raphson recurrence formula is:

$$u_k = u_{k-1} - \frac{u_{k-1}^2 - a}{2u_{k-1}}$$

$$= u_{k-1} - \tfrac{1}{2}u_{k-1} + \frac{1}{2}\frac{a}{u_{k-1}}$$

$$= \frac{1}{2}\left(u_{k-1} + \frac{a}{u_{k-1}}\right).$$

This is *Newton's Formula* for calculating the square root of a.

Section 5.4

Exercise 1

The quadratic Taylor approximation to the sine function about 0 gives

$$\sin\left(\frac{\pi}{10}\right) \simeq 1 \times \left(\frac{\pi}{10}\right) + \frac{1}{2} \times 0 \times \left(\frac{\pi}{10}\right)^2$$

$$= \frac{\pi}{10}$$

$$= 0.3142 \text{ to 4 decimal places.}$$

The quadratic Taylor approximation with $a = 0$ is in this case the same as the tangent approximation with $a = 0$, and is about twice as accurate as the tangent approximation with $a = \dfrac{\pi}{6}$ considered in Exercise 5.1.1

(the error is about 1.5% instead of 3%).

Section 5.5

Exercise 1

We have

$$c(x) = b_0 + b_1(x - a) + b_2(x - a)^2 + b_3(x - a)^3$$

$$c'(a) = b_1 + 2b_2(x - a) + 3b_3(x - a)^2$$

$$c''(x) = 2b_2 + 6b_3(x - a)$$

$$c'''(x) = 6b_3.$$

Therefore

$$c(a) = b_0 \quad \text{(and we are given that } c(a) = f(a))$$

$$c'(a) = b_1 \quad \text{(and we are given that } c'(a) = f'(a))$$

$$c''(a) = 2b_2 \quad \text{(and we are given that } c''(a) = f''(a))$$

$$c'''(a) = 6b_3 \quad \text{(and we are given that } c'''(a) = f'''(a)).$$

Thus, $b_0 = f(a)$, $b_1 = f'(a)$, $b_2 = \frac{1}{2}f''(a)$, $b_3 = \frac{1}{6}f'''(a)$, and so the formula for c is

$$c(x) = f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2 + \tfrac{1}{6}f'''(a)(x - a)^3.$$

Exercise 2

$$D \cos(x) = -\sin x \quad \text{gives} \quad D \cos(0) = \quad 0;$$

$$D^2 \cos(x) = -\cos x \quad \text{gives} \quad D^2 \cos(0) = -1;$$

$$D^3 \cos(x) = \quad \sin x \quad \text{gives} \quad D^3 \cos(0) = \quad 0;$$

$$D^4 \cos(x) = \quad \cos x \quad \text{gives} \quad D^4 \cos(0) = \quad 1;$$

and so on.

Therefore, the Maclaurin approximation to the cosine function contains only *even* powers of x; so for any positive integer n the Maclaurin polynomial approximations of degrees $2n$ and $2n + 1$ are the same, and are given by

$$\cos x \simeq 1 - \frac{1}{2}x^2 + \frac{1}{4!}x^4 + \cdots + (-1)^n \frac{1}{(2n)!}x^{2n}.$$

The approximation of degree 0 (or 1) to $\cos(0.3)$ is therefore

$$\cos(0.3) \simeq 1.$$

The approximation of degree 2 (or 3) is

$$\cos(0.3) \simeq 1 - (\tfrac{1}{2} \times 0.09) = 0.955.$$

The approximation of degree 4 (or 5) is

$$\cos(0.3) \simeq 1 - (\tfrac{1}{2} \times 0.09) + (\tfrac{1}{24} \times 0.0081) = 0.9553$$

which agrees with the true value to 4 decimal places.

Exercise 3

Let

$$f: x \longmapsto (1 - x)^s \qquad\qquad (x \in R, x \neq 1);$$

then

$$Df: x \longmapsto -s(1 - x)^{s-1} \qquad\qquad (x \in R, x \neq 1),$$

$$D^2f: x \longmapsto s(s - 1)(1 - x)^{s-2} \qquad (x \in R, x \neq 1),$$

and, for any $n \in Z^+$,

$$D^n f: x \longmapsto (-1)^n s(s - 1) \ldots (s - n + 1)(1 - x)^{s-n}$$

$$(x \in R, x \neq 1).$$

The general Maclaurin approximation of degree n is

$$(1 - x)^s \simeq 1 - sx + \frac{s(s - 1)}{1 \times 2}x^2 - \frac{s(s - 1)(s - 2)}{1 \times 2 \times 3}x^3 + \cdots$$

$$\cdots + (-1)^n \frac{s(s - 1) \times \cdots \times (s - n + 1)}{1 \times 2 \times \cdots \times n}x^n.$$

When s is a positive integer, the sth coefficient is $(-1)^s$, and each coefficient thereafter has a zero in the numerator, and is therefore equal to zero. You should recognize the above polynomial as the binomial expansion for $(1 - x)^s$, which gives the *exact* value of $(1 - x)^s$ when s is a positive integer.

When $s = -1$, the situation is vastly different. The general Maclaurin approximation of degree n is now

$$(1 - x)^{-1} \simeq 1 - (-1)x + \frac{(-1) \times (-2)}{1 \times 2}x^2$$

$$- \frac{(-1) \times (-2) \times (-3)}{1 \times 2 \times 3}x^3 + \cdots$$

$$\cdots + (-1)^n \frac{(-1) \times (-2) \times \cdots \times (-n)}{1 \times 2 \times \cdots \times n}x^n$$

$$= 1 + x + x^2 + \cdots + x^n.$$

In this case, since s is not a positive integer, there is no exact polynomial expression for $(1 - x)^s$, and so there is no "final" polynomial in the sequence of Maclaurin approximations.

(i) When $x = 0.1$,

$$(1 - x)^{-1} = \frac{1}{0.9} = \frac{10}{9} = 1.111 \ldots$$

The first approximation is   1.1;

the second approximation is   1.11;

the third approximation is   1.111;

etc.

(ii) When $x = 10$,

$$(1 - x)^{-1} = -\tfrac{1}{9} \qquad = -0.111 \ldots$$

The first "approximation" is   11;

the second "approximation" is   111;

the third "approximation" is   1111;

etc.

Section 5.6

Exercise 1

Taylor's Theorem tells us that the magnitude of the error, which equals the magnitude of the correction, cannot exceed

$$\tfrac{1}{2}Bx^2,$$

where $|\exp t| \leqslant B \, (t \in [x, 0])$.

Since $\exp t$ increases with t, and x is negative, the largest value of $\exp t$ for $t \in [x, 0]$ is $\exp 0 = 1$; so we may take $B = 1$. The magnitude of the error in the approximation,

$$\exp x \simeq 1 + x,$$

is therefore at most $\tfrac{1}{2}x^2$. In the particular case when $x = -0.2$ we have

$$\exp(-0.2) \simeq 1 - 0.2 = 0.8,$$

with a maximum error of $\tfrac{1}{2}(-0.2)^2 = 0.02$, from which it follows that

$$\exp(-0.2) \in [0.8 - 0.02, 0.8 + 0.02],$$

that is, $\exp(-0.2) \in [0.78, 0.82]$.

Exercise 2

The tangent approximation we are using is

$$\sin x \simeq \sin 0 + x \sin' 0$$
$$= x$$

Taylor's Theorem tells us that the magnitude of the error cannot exceed

$$\tfrac{1}{2}Bx^2,$$

where $|\sin'' t| \leqslant B \, (t \in [0, x])$.

Since $\sin' = \cos$ and $\cos' = -\sin$, we have $\sin'' = -\sin$, and so the condition on B reduces to

$$|\sin t| \leqslant B \qquad (t \in [0, x]).$$

For the case when $x = \dfrac{\pi}{10}$, we require a number B such that

$$|\sin t| \leqslant B \qquad \left(t \in \left[0, \frac{\pi}{10}\right]\right).$$

For a quick estimate of the maximum error we may use the fact that $\sin t$ always lies in the range $[-1, 1]$, and take $B = 1$. We thus obtain the following value for the maximum error:

$$\frac{1}{2} \times 1 \times \left(\frac{\pi}{10}\right)^2 = \frac{\pi^2}{200} \simeq \frac{9.9}{200} = 0.05,$$

which is a perfectly satisfactory answer to the question.

Alternatively, we can do a little more work and get the "best" (that is, the smallest possible) value of B.
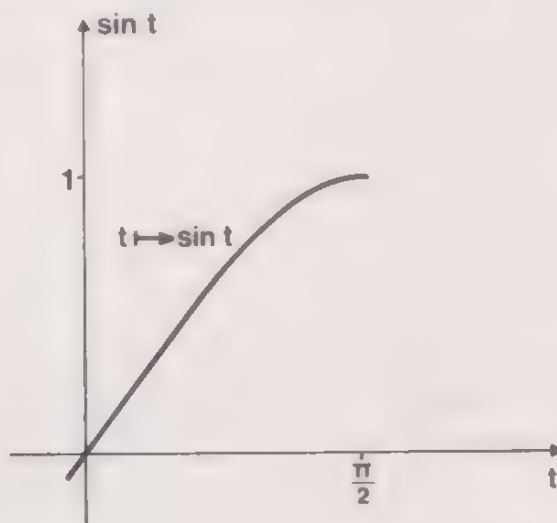
Since $\sin t$ increases with t in the interval $\left[0, \dfrac{\pi}{2}\right]$, its largest value in $\left[0, \dfrac{\pi}{10}\right]$ is $\sin \dfrac{\pi}{10} = 0.3090$. This gives a maximum error:

$$\tfrac{1}{2} \times 0.3090 \times \left(\frac{\pi}{10}\right)^2 \simeq 0.3090 \times 0.05 = 0.015.$$

(This is, of course, only of theoretical interest. Here we are discussing the accuracy of Taylor's approximation, but in a practical case we might

want to *calculate* $\sin\left(\dfrac{\pi}{10}\right)$ *using* Taylor's approximation, and then we

could not use our calculated value of $\sin\left(\dfrac{\pi}{10}\right)$ to obtain an error!) The

magnitude of the actual error is

$$\left|\sin\left(\frac{\pi}{10}\right) - \frac{\pi}{10}\right| = |0.3142 - 0.3090| = 0.0052.$$

Taylor's Theorem over-estimates the error by a factor of about 3 when the "best" value of B is used.

Section 5.7

Exercise 1

The required Maclaurin polynomials are:

$$\begin{aligned}
&\text{degree 1:} && \cos x \simeq 1 \\
&\text{degrees 2, 3:} && \cos x \simeq 1 - \tfrac{1}{2}x^2 \\
&\text{degree 4:} && \cos x \simeq 1 - \tfrac{1}{2}x^2 + \tfrac{1}{24}x^4
\end{aligned}$$

Taylor's Theorem tells us that

$$|C_n(x)| \leqslant \frac{1}{(n+1)!}B_{n+1}|x|^{n+1},$$

provided $|\cos^{(n+1)} t| \leqslant B_{n+1}$ $(t \in [0, x])$.

Since all the derived functions of cos are $\pm\cos$ or $\pm\sin$, B_n can be taken as 1 for all n: so, with $x = 2$, our problem is to find an n such that

$$\frac{1}{(n+1)!}2^{n+1} \leqslant \tfrac{1}{2} \times 10^{-2}, \quad \text{to ensure that } |C_n(x)| \leqslant \tfrac{1}{2} \times 10^{-2}.$$

Trying successive values of n, and using the table of factorials, we obtain:

$$\frac{1}{7!}2^7 = \frac{128}{5\,040} \simeq \frac{1}{50} > \frac{1}{2} \times 10^{-2}$$

$$\frac{1}{8!}2^8 = \frac{256}{40\,320} \simeq \frac{1}{160} > \frac{1}{2} \times 10^{-2}$$

$$\frac{1}{9!}2^9 = \frac{512}{362\,880} < \frac{600}{360\,000} = \frac{1}{600} < \frac{1}{2} \times 10^{-2}$$

Thus the conditions of the problem are satisfied with $n = 8$. You may have chosen a value of n larger than 8. This is also correct: it gives an even smaller maximum error.

The Maclaurin polynomial approximation of degree 8 for $\cos x$ is

$$\cos x \simeq 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \frac{1}{8!}x^8.$$

Section 5.8

Exercise 1

The formula given in Example 1 gives, for the kth partial sum,

$$S_k = 1 + r + r^2 + \cdots + r^{k-1} = \frac{1 - r^k}{1 - r} \qquad (r \in R, r \neq 1).$$

We are interested in the behaviour of this expression for large k. This depends on the value of r, and so there are several cases to consider.

(i) If $|r| < 1$, then $\lim_{k \text{ large}} r^k = 0$, and so

$$\lim_{k \text{ large}} \frac{1 - r^k}{1 - r} = \frac{1}{1 - r}.$$

In this case the series converges and its sum is $\dfrac{1}{1 - r}$.

(ii) If $r = 1$, then the formula for S_k does not apply; we see that $S_k = k$, and so the series diverges.

(iii) If $|r| > 1$, then $|r^k|$ increases with k, without any bound, and so the series diverges.

(iv) If $r = -1$, we have the series

$$1 - 1 + 1 - 1 + 1 \ldots$$

which diverges.

Exercise 2

The kth partial sum, S_k, of the infinite series is the $(k - 1)$th degree Taylor approximation to $f(x)$ about $x = a$. Thus we have

$$f(x) \simeq S_k$$

with correction $C_k(x)$; or, in other words,

$$f(x) = S_k + C_k(x).$$

Thus, if we are given that $\lim\limits_{k \text{ large}} C_k(x) = 0$, it follows that

$$f(x) = \lim\limits_{k \text{ large}} S_k.$$

Consequently, the infinite series converges and its sum is $f(x)$.

In the cases when $\lim\limits_{k \text{ large}} C_k(x)$ is either non-existent or different from zero, the conclusion is that the series either diverges or converges to a limit different from $f(x)$.

Exercise 3

The series

$$1 - x^2 + x^4 - x^6 + \cdots$$

is a geometric series with common ratio $(-x^2)$, and therefore has the sum

$$\frac{1}{1 - (-x^2)} = \frac{1}{1 + x^2},$$

whenever $|-x^2| = x^2 < 1$; that is, whenever $|x| < 1$. The series diverges if $|x| \geqslant 1$. The partial sums of this series are the polynomials

$$1 - x^2, \qquad 1 - x^2 + x^4, \qquad 1 - x^2 + x^4 - x^6, \ldots$$

which accordingly form a convergent sequence of approximations to $\frac{1}{1 + x^2}$ if $|x| < 1$, but not otherwise.

Exercise 4

Since the sequence of successive approximations obtained in the preceding exercise converges to $\frac{1}{1 + x^2}$ for all x in the interval of integration, with the single exception of the end-point 1, it is reasonable to guess that by successively substituting these polynomials for $\frac{1}{1 + x^2}$ in the integral we shall obtain a sequence of successive approximations to $\frac{\pi}{4}$. This sequence is:

$$\int_0^1 (x \longmapsto 1 - x^2) = \left[x \longmapsto x - \frac{1}{3}x^3 \right]_0^1 = 1 - \frac{1}{3}$$

$$\int_0^1 (x \longmapsto 1 - x^2 + x^4) = \left[x \longmapsto x - \frac{1}{3}x^3 + \frac{1}{5}x^5 \right]_0^1 = 1 - \frac{1}{3} + \frac{1}{5}$$

and so on: the nth approximation in the sequence is

$$\frac{\pi}{4} \simeq 1 - \frac{1}{3} + \frac{1}{5} - \cdots + (-1)^n \frac{1}{2n+1}.$$

The corresponding infinite series is

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots.$$

Section 5.9

Exercise 1

The equation is most conveniently put in the form $f(x) = 0$ by taking

$$f(x) = x - \sin x - \tfrac{2}{3}\pi.$$

Then $f'(x) = 1 - \cos x$, and the Newton–Raphson recurrence formula is:

$$u_k = u_{k-1} - \frac{u_{k-1} - \sin(u_{k-1}) - \tfrac{2}{3}\pi}{1 - \cos(u_{k-1})}.$$

The value of u_1 is not all that critical, but it is sensible to try to get it near to the solution. We have chosen $u_1 = 2$; you may well have chosen some other value, but you should get the same final result. Working to 3 decimal places, we obtain

$$u_1 = 2$$

$$u_2 = 2.709$$

$$u_3 = 2.607$$

$$u_4 = 2.605$$

$$u_5 = 2.605.$$

Exercise 2

$$\exp(1) = \exp'(1) = \exp''(1) = 2.72$$

to 2 decimal places.

Therefore, using the quadratic Taylor approximation, we find:

$$\exp(1.2) \simeq 2.72(1 + 0.2 + \tfrac{1}{2}(0.2)^2)$$

$$= 2.72 \times 1.22$$

$$= 3.32 \text{ to 2 decimal places}$$

which agrees with the true value of $\exp(1.2)$ to 2 decimal places.

Exercise 3

Since the nth derivative of exp at x is $\exp(x)$ for all n, the general Maclaurin approximation of degree n for $\exp(x)$ is simply

$$\exp(x) \simeq \exp(0) + x\exp(0) + \tfrac{1}{2}x^2\exp(0) + \cdots + \frac{1}{n!}x^n\exp(0)$$

$$= 1 + x + \tfrac{1}{2}x^2 + \cdots + \frac{1}{n!}x^n.$$

The first-degree approximation to $\exp(0.1)$ is

$$\exp(0.1) \simeq 1 + 0.1 = 1.1.$$

The second-degree approximation is

$$\exp(0.1) \simeq 1 + 0.1 + 0.005 = 1.105.$$

The third-degree approximation is again

$$\exp(0.1) \simeq 1.105 \text{ to 3 decimal places,}$$

and the fourth and higher degree approximations also give 1.105. Thus the Maclaurin approximations for $\exp(0.1)$ converge much more quickly than the calculation directly from the definition — the second-degree approximation is already correct to 3 decimal places.

Exercise 4

$$\exp(1) = \exp'(1) = \exp''(1) = 2.72 \text{ (to 2 decimal places).}$$

Therefore we have the following estimates (to 2 decimal places) for the correction to the tangent approximation at 1 for $x = 0.8, 0.9, 1.1, 1.2$:

$$C_1(0.8) \simeq \tfrac{1}{2} \times 2.72 \times (-0.2)^2$$

$$= 0.05$$

$$C_1(0.9) \simeq \tfrac{1}{2} \times 2.72 \times (-0.1)^2$$

$$= 0.01$$

$$C_1(1.1) \simeq 0.01$$

$$C_1(1.2) \simeq 0.05$$

CHAPTER 6   FIRST ORDER DIFFERENTIAL EQUATIONS

6.0   Introduction

In Volume I, Chapter 8 we saw that the derivative of a real function Q represents the rate of change of the image value under Q, or the slope of the (tangent to the) pictorial graph of Q; that is, $Q'(t)$, $(t \in R)$, is the rate of change of $Q(t)$ at t. Given the function Q, we can derive Q' by using one of the standard rules. For example, if

$$Q : t \longmapsto e^t - 2t^3 \qquad (t \in R),$$

then

$$Q' : t \longmapsto e^t - 6t^2 \qquad (t \in R).$$

This is straightforward.

The problem is appreciably harder when we are given the derived function Q' and are asked to determine Q. This is the problem of *integration* and we have developed some tools to cope with this (Volume I, Chapters 7 and 9, and Chapter 3 in this volume). Essentially, there are two main points. In the first place, the differentiation operator D is a function, but the reverse of D is not a function. We coped with this by introducing a constant function and obtained the images of Q' under the reverse mapping in the form

any primitive + constant function.

For example, if we are given

$$Q' : t \longmapsto 2t \qquad (t \in R)$$

then

$$Q : t \longmapsto t^2 + c \qquad (t \in R),$$

where c is a real number, and for each choice of c we get an image of Q' under the integration mapping. The second important point is that very often we could not find a simple expression for the integral (or, even if we could, the labour involved was prohibitive).

For example, if

$$Q' : t \longmapsto \frac{1}{\sqrt{t^3 + 1}} \qquad (t \in R^+)$$

then

$$Q : t \longmapsto ?$$

In this chapter we are going to make our integration process still more difficult; not because it is "good for the soul", but because of the wide field of application of the development. We are on the threshold of the subject known as *differential equations*, which has a considerable literature associated with it and on which much research is still undertaken today.

A *differential equation* is an equation involving an unknown real function and its derived functions, and possibly other known functions. Thus if f is a real function,

$$f' + 2f = O$$

(where O is the function $x \longmapsto 0 \ (x \in R)$) is a differential equation, as is

$$f' + (x \longmapsto x^2) \times f = x \longmapsto \sin x.$$

The simplest form of differential equation is an equation of the form

$$f' = g$$

where g is a known real function; its solution presents us with precisely the integration problem we have discussed previously. We shall discuss more fully in section 2 what we mean by the *solution* of a differential equation, but we note here that the solution of such an equation is a *function* (or *set of functions*).

For example, one element of the solution set of the differential equation

$$f' + 2f = O$$

is the function

$$f : x \longmapsto -2 \exp(-2x) \qquad (x \in R).$$

Note that a relation between functions can be written in terms of the images under the functions. For example,

$$f' + 2f = O$$

is equivalent to

$$f'(x) + 2f(x) = 0,$$

for all x in the domain of f'. We shall refer to the "image form" of a differential equation as a differential equation too; in this case it is frequently convenient to use the Leibniz notation for the derivative (see General Note below).

Differential equations arise directly from many basic physical laws and are therefore fundamental to the study of considerable parts of science and engineering. For example, one of Newton's laws states that

force applied = mass × acceleration.

Acceleration is the rate of change of velocity, and velocity depends on time. If, for instance, the force applied itself depends on the velocity, as it does in the car engine, then we can write the above equation in terms of a "position" function and its first and second derived functions, and hence obtain a differential equation. Differential equations do not arise solely from this kind of background: they also arise from such diverse sources as architecture, biology and economics. In this chapter we are going to solve only a simple form of differential equation involving Q, Q' and known functions.

Our plan is to concentrate mainly on a few closely related differential equations. In section 1 we shall do the modelling; in other words, we shall develop some differential equations from plausible physical situations. In section 2 we shall discuss some basic ideas concerning the solution of differential equations. We shall then investigate graphical solutions of differential equations, since the pictorial approach will give us a good idea of how solutions behave. In sections 4 and 5 we introduce some which might be called *formula* methods of solution of differential equations. At this particular stage the objective of a formula method is to rearrange the differential equation in such a way that we can use the methods of integration introduced earlier.

General Note

From now on we shall use, as appropriate and convenient, any of the calculus notations mentioned previously in the course. For instance, if Q is a real function, such that

$$Q : t \longmapsto Q(t) \qquad (t \in R)$$

and we write $q = Q(t)$, then we write the derived function as Q', when the "function" notation suits us, or we express the derivative as

$$Q'(t) = \frac{dq}{dt},$$

using the "function image" notation or the Leibniz notation. You should by now be reasonably familiar with the function notation as we have used it throughout, and also with the distinction between the function Q and the variables t and q, where t is a general element of the domain of Q,

and q is a general element of the image set. Our objective in using these various notations is to enable you not only to read this book, but also to read any other texts that you may come across now and in further studies.

6.1   Population Growth

In this section we take up a problem of population growth considered previously in section 5.1 of Volume I, and look at the increasingly sophisticated mathematical models we can design.

Let $q = Q(t)$ represent (at any time $t \in R$) the population (in thousands) of a particular species, which we shall take to be human beings, but could equally well be viruses, locusts, fish or birds.

Since q is measured in thousands, it can really take only certain rational values corresponding to a whole number of human beings, e.g. 285.632. This brings us to our first modification in the process of formulating our mathematical model. To talk about growth, or rate of change, we want to talk about the *derived function*, and the derived function was certainly not defined for functions with this subset of the rationals as codomain since the limiting procedure we adopted in the definition then becomes meaningless. This looks like a full stop: but what we do is to take the co-domain of Q to be R^+ in order to make our mathematical model. That is, we assume that the variable q can take *any* positive real values to enable us to use the powerful tools of calculus.

To determine how the population will change with time we need to intro-duce functions to represent the number (in thousands) of births and deaths per year. We shall call these B and M respectively. The number of births (deaths) usually depends on the size of the population, $Q(t)$, which itself depends on the time at which it is measured. Thus the composite function $B \circ Q$ will tell us how the number of births per year depends on the time.

The domain of each of the functions B and M is the subset of the reals which represents the population in thousands, and the codomain is a sub-set of the reals which represents the number of births (deaths) in thousands per year. We shall measure t in years. The rate of growth of population is then given by

$$DQ = B \circ Q - M \circ Q$$

or

$$Q'(t) = B(Q(t)) - M(Q(t)) = B(q) - M(q)$$

The simplest assumption to make is that the number of births and the number of deaths per year are both constant, i.e.

$$B : q \longmapsto b_0 \qquad (q \in R^+),$$

$$M : q \longmapsto m_0 \qquad (q \in R^+),$$

where b_0 and m_0 are known numbers.

Thus we get

$$Q'(t) = b_0 - m_0 = k_0, \text{ say.}$$

It is not our purpose to *solve* differential equations in this section, although this one has obvious solutions. All we are illustrating here is how differential equations arise. It is unrealistic to imagine that the number of births per year will remain constant; for example, if the population doubles it is likely that the number of births also doubles. To translate this into mathematical terms is to require that the function B be of the form

$$B : q \longmapsto b_1 q \qquad (q \in R^+),$$

where b_1 is a positive number. That is, the number of births is proportional to the number in the population.

Similarly, it is reasonable to assume that the number of deaths is proportional to the population, i.e.

$$M : q \longmapsto m_1 q \qquad (q \in R^+),$$

where m_1 is a positive number. The differential equation is now

$$Q'(t) = b_1 Q(t) - m_1 Q(t) = k_1 Q(t)$$

or

$$\frac{dq}{dt} = k_1 q,$$

where $k_1 = b_1 - m_1$.

Now we introduce a further refinement to our model. As the population increases, so the available food supplies may become depleted. With a greater possibility of disease and disasters claiming a higher number of deaths, we would expect that the death rate would increase with increasing population.

We could perhaps represent this mathematically by

$$M : q \longmapsto m_1 q + m_2 q^2 \qquad (q \in R^+)$$

where m_2 is a positive number. For instance, if we take $m_1 = 0.02$ and $m_2 = 0.004$, we would have

$$M : q \longmapsto 0.02q + 0.004q^2 \qquad (q \in R^+);$$

the graph of M is depicted in the figure.
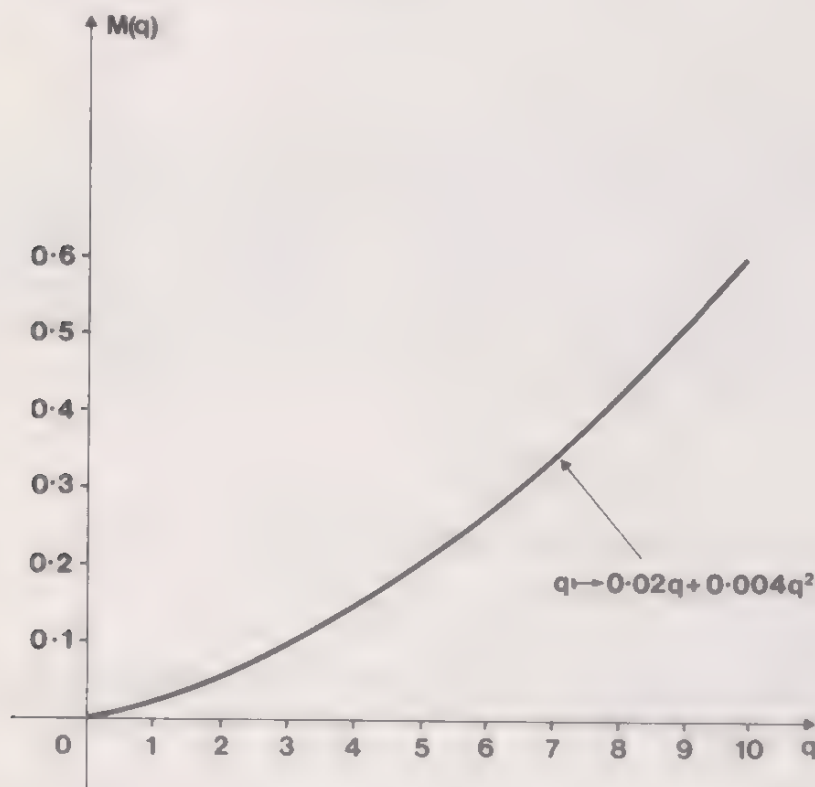
This would imply that with a population of 1 000, i.e. $q = 1$, the death rate would be 24 per thousand per year, whereas if the same population were to increase to 10 000, i.e. $q = 10$, the death rate would increase to $\dfrac{0.2 + 0.4}{10} = 0.06$, i.e. 60 per thousand per year. The differential equation would now be $= k_1 Q(t) - m_2 (Q(t))^2$

or $\qquad \dfrac{dq}{dt} = k_1 q - m_2 q^2,$

where $k_1 = b_1 - m_1$.

From now on in this text we shall write the "image form" of a differential equation in terms of either $Q(t)$ or q, rather than continue to write both versions. The first form, explicitly in terms of the images, is useful, since it tells us that we are looking for a *function* as a solution rather than a number. The second form, using the Leibniz form of the derivative, is convenient

because it is more concise. We shall use either form depending on which we feel is more appropriate in the context.

It is interesting to notice that we can obtain some information from a differential equation even without solving it. We can write our differential equation in the form

$$\frac{dq}{dt} = m_2 q\left(\frac{k_1}{m_2} - q\right).$$

Assuming $k_1 > 0$ $\left(\text{if it were not, } \frac{dq}{dt} \text{ would be negative and the population}\right.$ would decrease until the model became inapplicable or the population became zero$\Big)$ we see that, provided

$$q < \frac{k_1}{m_2},$$

$\frac{dq}{dt}$ is positive, implying increasing population. (Remember that $q = Q(t)$, i.e. q is a *number*.) If $q = \frac{k_1}{m_2}$, then the population is static, and if $q > \frac{k_1}{m_2}$, the derivative is negative and the population is decreasing. This tells us that $\frac{k_1}{m_2}$ is the only stable population.

This leads us to the final refinement we propose to make here. This stable population which can be supported is likely to increase with time, due to improvements in medicine, agricultural technology, etc. Let us suppose that this stable population increases linearly with time and that we replace $\frac{k_1}{m_2}$ by $(k_2 t + k_3)$, where k_2 and k_3 are positive numbers. The differential equation becomes

$$\frac{dq}{dt} = m_2 q (k_2 t + k_3 - q).$$

Notice that the right-hand side now involves the two variables q and t. These variables are customarily given special names; t, the domain variable of the function Q we wish to determine is called the independent variable; q is the dependent variable, since it is the variable in the codomain of the function Q we wish to determine.

Notice also that we have made various suppositions without any real

justification. What we have is a very *tentative* model which, on analysis, may or may not fit the facts. In a real situation we would need to test its consequences against known data before we could attempt to make any predictions from it.

The various differential equations in this section are all differential equations of the *first order*. Order is determined by the highest derivative present. In other words, if a differential equation contained $\dfrac{d^n q}{dt^n}$ or $Q^n(t)$ and no higher derivatives, we would say it was *n*th *order* or *of order n*. For example, the equation

$$\left(\frac{d^3 q}{dt^3}\right)^2 = 9$$

is third order.

6.2 Basic Ideas about Solutions

Consider the familiar problem of solving the quadratic equation

$$t^2 - 5t + 4 = 0 \qquad (t \in R).$$

The solution set (see Volume I, page 31) of this equation, namely

$$\{t : t^2 - 5t + 4 = 0, t \in R\},$$

contains two members and is

$$\{4, 1\}.$$

Other solution sets in R for quadratic equations may contain two, one (in the case of two equal roots) or no members.

Exercise 1

Write down examples of three quadratic equations whose solution sets contain two, one and no real members respectively.

Exercise 2

Describe, in an explicit form, the members of the solution set of the equation

$$\sin t = 0 \qquad (t \in R)$$

Can you count how many members there are?

The solution set of a differential equation may be written in a similar way to that of an ordinary equation. Consider the differential equation

$$Q'(t) = k_0 \qquad (t \in R)$$

or

$$Q'(t) - k_0 = 0$$

The set of all functions which satisfy this differential equation can be denoted by

$$\{Q : Q'(t) - k_0 = 0 \qquad (t \in R)\}.$$

In this case each solution to the differential equation is simply a primitive function (indefinite integral) of $t \longmapsto k_0$, that is

$$t \longmapsto k_0 t + c \qquad (t \in R);$$

c is a constant of integration. Thus another (explicit) form of the solution set is

$$\{Q : Q = t \longmapsto k_0 t + c \qquad (t \in R), c \in R\},$$

which is often more concisely written in terms of images as

$$\{Q : Q(t) = k_0 t + c \qquad (t \in R), c \in R\}.$$

This set can also be written as a relation between variables:

$$\{Q : q = k_0 t + c \qquad (t \in R), c \in R\}.$$

We see again here the important difference between the solutions of ordinary equations defined on sets of numbers and the solutions of differential equations which are defined on sets of functions. In the former case the solution set has *real numbers* as members whilst in the latter case it has *real functions* as members. We can observe this visually by considering the graphical representation of the solution set of the equation

$$\sin t = 0 \qquad (t \in R),$$

and comparing it with the graphical representation of the solution set of the differential equation

$$Q'(t) - 1 = 0 \qquad (t \in R)$$

We know from Exercise 2 that the solution set of

$$\sin t = 0 \qquad (t \in R)$$

is the *set of numbers*

$$\{n\pi : n \in Z\}.$$

Graphically, this can be represented as the intersection of the straight line which is the pictorial graph of

$$t \longmapsto 0$$

and the curve which is the pictorial graph of

$$t \longmapsto \sin t.$$

The points of intersection have co-ordinates $(n\pi, 0)$, $n \in Z$.
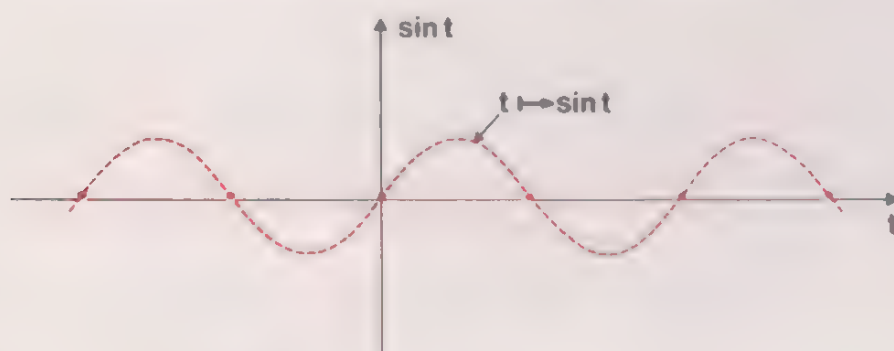
The solution set of the differential equation

$$Q'(t) - 1 = 0 \qquad (t \in R)$$

may be represented graphically by the *family of curves*, or set of solution curves, given by the pictorial graphs of the functions

$$Q : t \longmapsto t + c \qquad (t \in R), c \in R.$$

The fact that there is a *family* arises from all the different values that the constant of integration c may take. We illustrate the cases $c = 2, 1, 0, -0.6, -1, -2$.
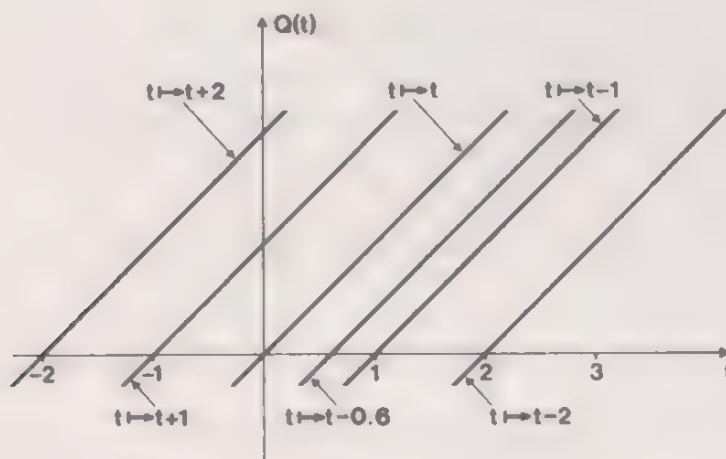
The whole plane is filled by curves (straight lines in this case). Through any point, one, and only one, curve of the family passes. Picking out one point in the plane picks out one particular curve, which is equivalent to picking out a particular constant of integration. For example, consider a population of fish which starts ($t = 0$) with a population of 2 000 and increases at the constant rate of 1 000 per year (difference between number of births and number of deaths).
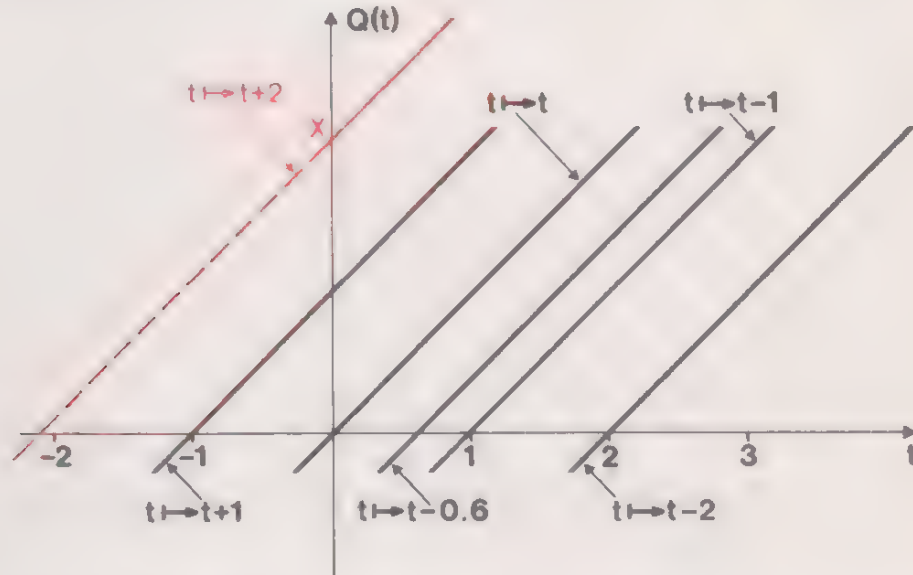
Then, using units of 1 000, we have an initial point $(0, 2)$ (marked with an X on the diagram), and the appropriate solution curve (marked in red on the diagram) is the graph of

$$Q : t \longmapsto t + 2 \qquad (t \in R)$$

since our general solution (i.e. $Q : t \longmapsto t + c$) implies

$$Q(0) = 0 + c,$$

i.e.

$$2 = 0 + c.$$

In a differential equation obtained as a mathematical model of a real situation, we usually know an initial assigned value, as in this example. Here, if the modelling is satisfactory, the one particular solution curve through the point

(initial time, initial population),

can be used to determine the future population. This process of picking a *particular curve* by the known *initial condition* is an important feature of the mathematical modelling in this situation.

Exercise 3

Sketch a typical subset of the family of curves which illustrates the solution set of the differential equation

$$Q'(t) - 2t = 0 \qquad (t \in R).$$

Indicate the particular solution curve which satisfies the initial condition $Q(1) = 2$. What is the numerical value of the constant of integration then?

Exercise 4

Rewrite

$$Q'(t) = 2t$$

in a form involving

(i) functions only,
(ii) variables only.

A differential equation, like an ordinary equation, may have no members in its solution set. Thus

$$\{f : (f'(t))^2 + (f(t))^2 + 1 = 0 \qquad (t \in R)\}$$

is an empty set, just as

$$\{t : t^2 - 2t + 2 = 0 \qquad (t \in R)\}$$

is empty. We have just seen that the number of solutions of $Q'(t) - 2t = 0$ is not finite. Also a differential equation can have a finite number of functions in its solution set. Thus

$$\{f : (f'(t))^2 + (f(t))^2 = 0 \qquad (t \in R)\}$$

has one member,

$$f : t \longmapsto 0 \qquad (t \in R),$$

just as

$$\{t : t^2 - 2t + 1 = 0 \qquad (t \in R)\}$$

has only one member.

Summary

Solutions of differential equations are functions. The pictorial graphs of these solutions form a family of solution curves. In our example we used

one initial condition (one point in the plane), to pick out one particular function (one particular solution curve).

6.3   Graphical Methods of Solution

We wrote down the solution of the equation

$$\sin t = 0$$

straight away, since we knew that

$$\sin(n\pi) = 0 \qquad (n \in Z),$$

and no other values of t in the domain R of the sine function are mapped to zero. But suppose that we have a less tractable equation of the form $f(t) = 0$ to solve: for example, the equation:

$$t - \sin t - \frac{2\pi}{3} = 0.$$

We can use a graphical method, first calculating the images of the function

$$t \longmapsto t - \sin t - \frac{2\pi}{3} \qquad (t \in R)$$

at a few selected points in the domain, and joining them up smoothly (on the correct assumption that the function involved is continuous). Then the value of t where the sketched curve crosses the x-axis will be the approximate solution we want.

Rough sketch of $t \longmapsto t - \sin t - \frac{2\pi}{3}$

169
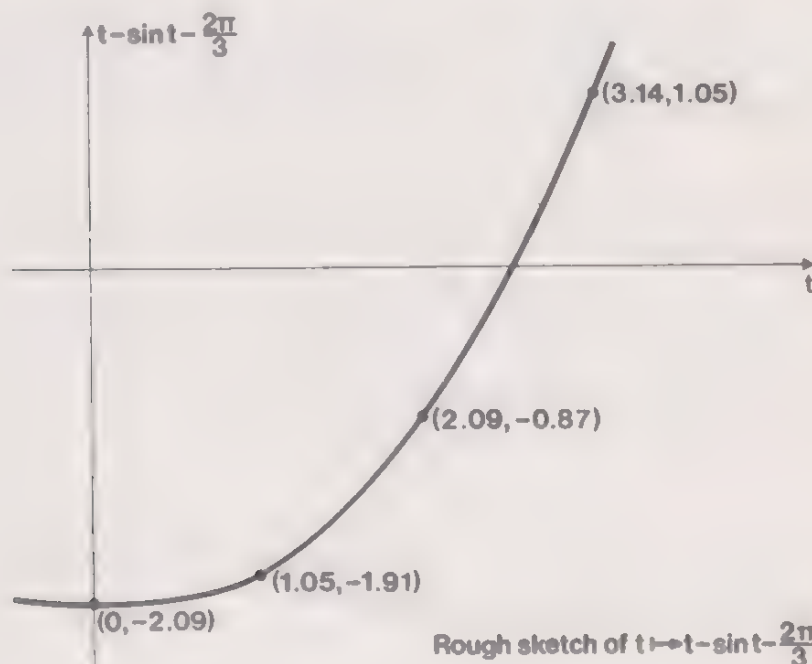
To improve the accuracy, we can magnify the portion of the graph in the neighbourhood of the approximate solution, by calculating new images of the function and drawing that part of the graph more precisely.

We can adopt a similar approach for the solution of first order differential equations; this approach will be useful for those that we cannot solve by the standard methods that we discuss in this text. The equation in the population growth example of section 6.1.

$$\frac{dq}{dt} = m_2 q(k_2 t + k_3 - q)$$

is one of this type. We shall however use the previous equation:

$$\frac{dq}{dt} = k_1 q - m_2 q^2$$

to illustrate the graphical method. Of course, to solve it graphically we must use actual numbers for k_1 and m_2. This illustrates the restrictive aspect of the graphical or numerical approach — not only do we have to specialize the problem in this way, but, in any case, we do not get a general expression for the solution. Suppose we choose $k_1 = 2$ and $m_2 = 1$. The differential equation is then

$$\frac{dq}{dt} = 2q - q^2$$

and we want to find the function, or functions, Q, such that $q = Q(t)$, which satisfy this equation. We shall be content to have such a function, or functions, specified graphically. We shall call the corresponding curves *solution curves*.

We know that $\frac{dq}{dt} = Q'(t)$ represents the slope of the graph of the function Q, that is, the direction in which the solution curve is pointing at any point (t, q). For our differential equation we can calculate the slope of a solution curve at any point; it is always given by

$$2q - q^2.$$

Thus at the point $(0, 1)$, representing a population of 1 000 at time $t = 0$, the slope of the particular solution curve is

$$2 \times 1 - 1^2 = 1$$

Since $2q - q^2$ is not explicitly dependent on t, the slope at any point $(t, 1)$ is always 1. In terms of the population growth which the differential equation represents, this means, as we would expect, that the rate of increase is explicitly dependent on the size of the population and not the time at which we are measuring it. In terms of the solution curves, this means that for $q \geqslant 0$ we have a solution curve through every point (t, q), and for each q the tangents to the solution curves at (t, q) are parallel for all t.

Let us plot slopes at a selection of points on an appropriate graph. Remember that the slope is the ratio:

$$\frac{\text{vertical distance}}{\text{horizontal distance}}.$$

Let us therefore fix a horizontal distance of, say, 4 small squares on the graph paper, and then draw the arrow for a particular point to indicate the direction of the solution curve there. For example, for the point $(1, 0.5)$ we need to draw an arrow to represent a slope of 0.75. The diagram indicates how this is done.
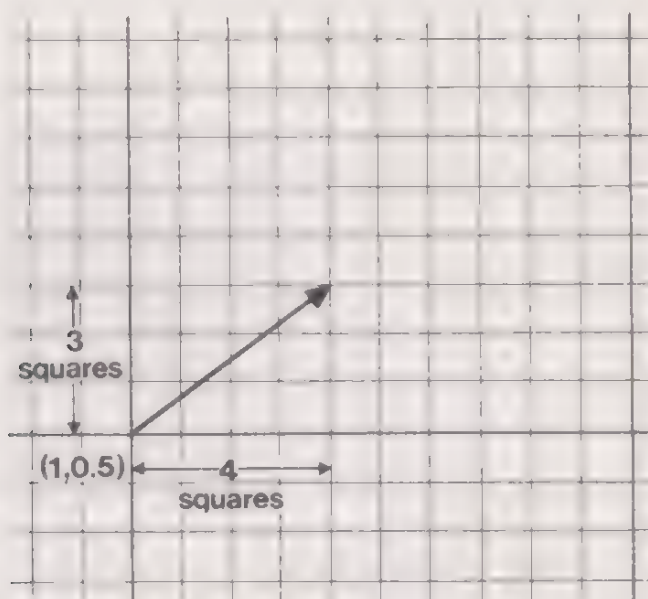
The actual numerical values of the slopes at some more selected points are displayed in the table below. As we have noted, the value of the slope does not depend on the value of t; so all the columns are the same. We have nevertheless set out a table, both to give you an idea of what happens in a more general case and to produce a record which is similar to the graphical picture.

q \ t	0	0.5	1	1.5	2	2.5
3	− 3	− 3	− 3	− 3	− 3	− 3
2.5	− 1.25	− 1.25	− 1.25	− 1.25	− 1.25	− 1.25
2	0	0	0	0	0	0
1.5	0.75	0.75	0.75	0.75	0.75	0.75
1	1	1	1	1	1	1
0.5	0.75	0.75	0.75	0.75	0.75	0.75
0	0	0	0	0	0	0

If we draw the corresponding arrows, we obtain the following diagram:
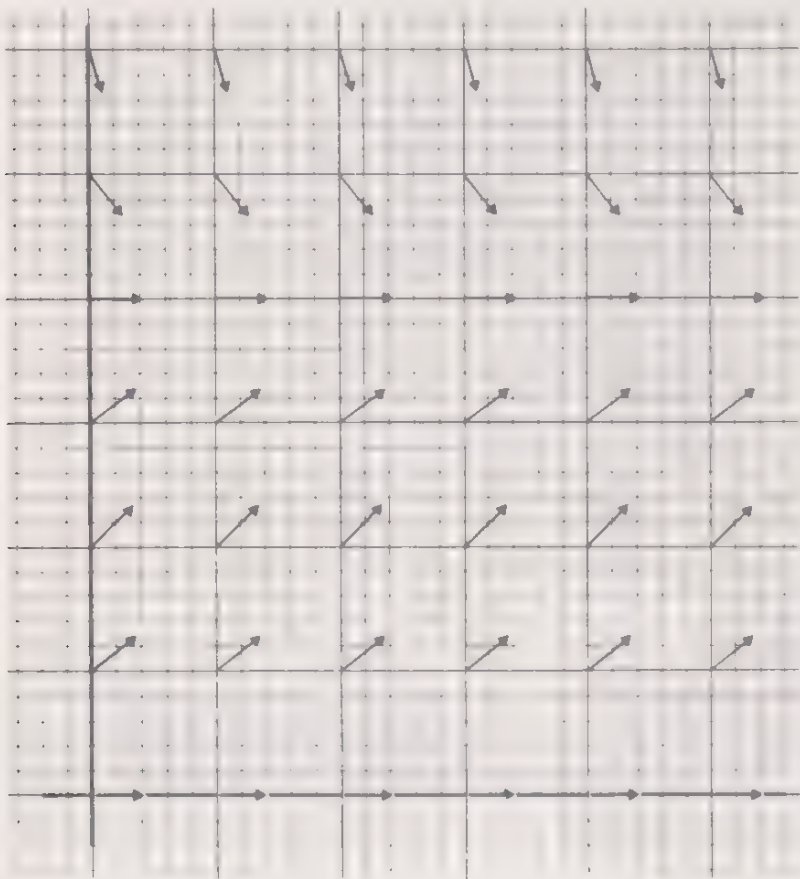
There are several points to note about the table and the diagram:

  (i) Since all the numbers in any one row are the same, wherever solution curves cross a given horizontal line they must all be pointing in the same direction.

 (ii) We are interested only in that portion of the plane for which $q \geqslant 0$ and $t \geqslant 0$ (called *the first quadrant*) since this is the portion which is meaningful in the context of this problem.

(iii) We could go on to calculate the slope at more and more points in the first quadrant and make the set of arrows as numerous and dense as we please (physical width and length of pencil marks allowing).

(iv) In the same sense that we sketch a curve through a few points, so too we can sketch a curve, or curves, using the arrows as guides. The following diagram shows a few typical solution curves. All we have to do to sketch a solution curve is to start from a given point and make sure that the direction of the curve is roughly parallel to all the arrows it passes close to.
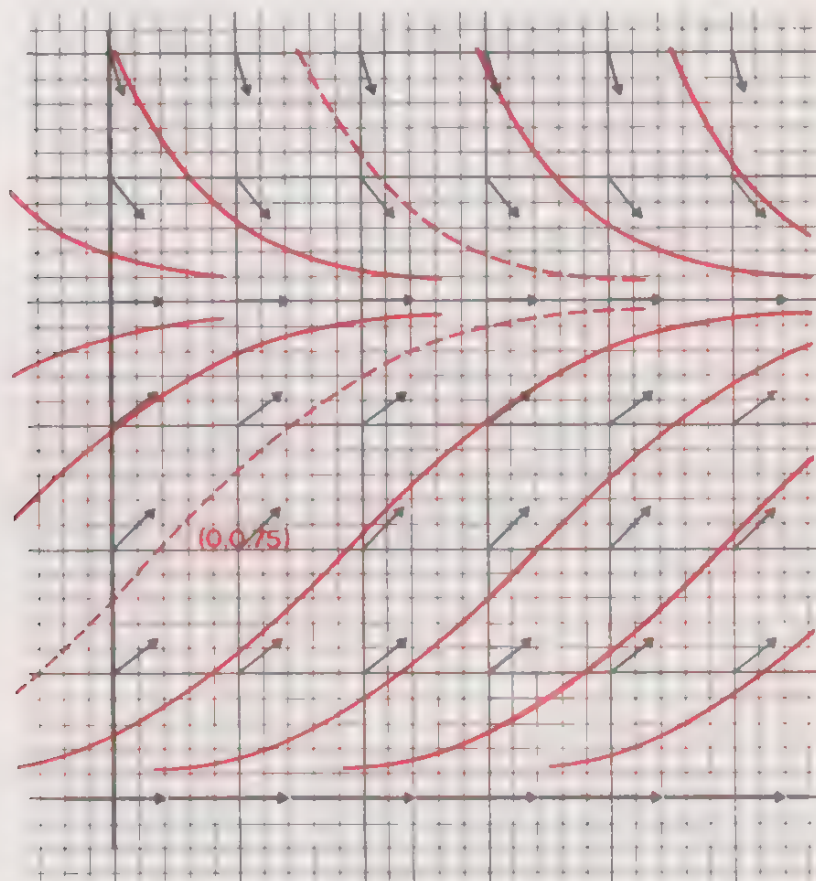
  (v) If we wish to improve the accuracy, we must use more arrows, just as we would use more points to draw an ordinary graph.

 (vi) What we have produced is an approximation to a selection of curves from the family of curves which represent the solution to the differen-

tial equation. We have thus, albeit laboriously, got the picture of what the family of solution curves look like. If you look at the set of arrows at a distance you get the impression of how the pattern "flows".

(vii) For our particular differential equation, we can see by substitution that $q = 0$ is a trivial solution: if the population is zero initially, mathematical and other reasoning tells us that the population will never be anything but zero. When $q > 0$, all the solution curves appear to approach the same numerical value 2 as t becomes large. 2 000 is thus the stable population figure. If there were more than 2 000 people originally they will reduce to this number: if less they will increase to this number. As we mentioned in section 6.1, we could have deduced this *particular* piece of information *without solving* the differential equation. For the derivative, $Q'(t)$, is zero when

$$2q - q^2 = 0,$$

that is when $q = 0$ (and we have dispensed with that) and when $q = 2$. The value $q = 2$ is never attained in our present model except in the special case when the population starts with 2 000.

(viii) To select the *particular* solution curve representing a particular case of interest, we need more information; e.g. an initial condition. Suppose we start with 750 in the population; that is, $q = 0.75$, and choose to measure time from that instant, so that $q = 0.75$ when $t = 0$. We start to draw our curve from this point and produce the approximate solution curve indicated by a red dotted line in the previous diagram. After 2 years we see (from our curve) that the population of 2 000 has almost been attained. After one year we estimate that there would be 1 700 in the population.

The graphical method works for all first order differential equations which we can put in the form

$$\frac{dq}{dt} = f(t, q),$$

with certain conditions on f. All we have to do is to evaluate the images under f at appropriate points in the plane. The conditions on f which ensure that a solution exists are too complicated for this text.

The graphical method can be a very useful method in cases which cannot be solved by any other means. It can be inaccurate (as can curve-sketching of any type) and it is certainly laborious, but one advantage is that we get a good qualitative idea of the shapes of all the solution curves, that is, the

solution set as a whole. With this "panoramic view" we can choose particular points of interest and find numerically the particular solution curves which pass through them.

Summary

In the graphical approach we use the geometric information given by the differential equation

$$\frac{dq}{dt} = f(t, q),$$

namely the slopes of the solution curves at all the points (t, q) of the domain of f, to sketch the family of solution curves.

6.4   Formula Method 1: Separation of Variables

In the next two sections we pick out from the repertoire of *formula methods* of solution two which will give exact solutions to first order differential equations of certain types. (The solutions are exact in the sense that we can specify precisely the set of functions which form the solution set, rather than give them in tabulated form as with a numerical method.) The usefulness of formula methods lies in the fact that they produce a more "general" solution than the graphical method, just as the solutions of the quadratic equation.

$$ax^2 + bx + c = 0$$

are written more generally as

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

than by saying that if $a = 1$, $b = -2$ and $c = 1$, then $x = 1$, because the general solution can be evaluated for many different numerical values of a, b, c.

Why discuss only two formula methods? The reason is that the two we have chosen are simple enough to grasp and learn to use and are representative of the exact methods available. It is not our intention to list all the possible "recipes" for solving equations — after all, we do not intend that this should be a cookery book. Rather, we hope that you will gain a general idea of what is involved in finding exact solutions of differential equations,

and will be able and prepared to learn other methods if and when you need them.

All the exact methods necessarily rely on the methods of integrating and differentiating functions we discussed earlier. We shall quote them explicitly as required.

The method discussed in this section is useful only when we can "disentangle" the variables in the equation.

We illustrate the method by an example. Consider the equation

$$Q'(t) = tQ(t) \qquad (t \in R).$$

We regard t and $Q(t) = q$ as variables, and we can "disentangle" them by rearranging the equation in the form

$$\frac{Q'(t)}{Q(t)} = t$$

provided that we exclude any t such that $Q(t) = 0$ from the domain of Q. Writing this in function (rather than image) form we get

$$\frac{1}{Q} \times DQ = t \longmapsto t \qquad (t \in R).$$

Since $\dfrac{1}{Q} \times DQ$ and $t \longmapsto t$ are equal functions, we can equate their primitives provided that we choose the constants of integration correctly. We therefore try to "integrate both sides".

We can integrate the right-hand side to get the set of primitive functions of the form $t \longmapsto \dfrac{t^2}{2} + c, c \in R$, but the integral of the left-hand side may not be so obvious. In Chapter 3, Section 2 we obtained the following formula for the primitive function of a composite function:

$$\int (g \circ k) \times Dk = \left(\int g \right) \circ k$$

Comparing this with our equation, we see that we need to choose $k = Q$ and $g \circ k = \dfrac{1}{Q}$, that is

$$g : t \longmapsto \frac{1}{t}$$

This function has a standard integral, viz

$$\int g = t \longmapsto \ln t + c_1 \qquad\qquad (t \in R^+)$$

and

$$\left(\int g\right) \circ k = t \longmapsto \ln Q(t) + c_1 \qquad (Q(t) \in R^+),$$

where c_1 is any real number. Notice how another condition has crept in here, that is, $Q(t) > 0$ (because the logarithm function has domain R^+). We could deal with $Q(t) < 0$ if it were of any interest, but we would have to do it separately. When using standard results, as we are here, we must always be careful to use them in the appropriate circumstances.

Putting the bits together, we have

$$t \longmapsto \ln Q(t) + c_1 = t \longmapsto \frac{t^2}{2} + c$$

This specifies the set of functions Q which belong to the solution set of the original equation, but in an inconvenient form: the dependent variable $q = Q(t)$ is not expressed explicitly in terms of the independent variable t. So we try to reorganize. First we use the simpler form in terms of images:

$$\ln q + c_1 = \frac{t^2}{2} + c$$

i.e.

$$\ln q = \frac{t^2}{2} + (c - c_1)$$

Since c and c_1 are any real numbers, $c - c_1$ is any real number. We can, therefore, either write $d \in R$ for $c - c_1$ or just drop the c_1. (In general, we do not need to introduce the constant of integration on both sides of an equation, since we know that if two functions are equal, their primitives differ by a constant.) So now we have

$$\ln q = \frac{t^2}{2} + d$$

Remembering that ln is a one-one function and exp is its inverse, we have

$$q = \exp\left(\frac{t^2}{2} + d\right)$$

We now have the dependent variable $q = Q(t)$ expressed explicitly in terms of the independent variable t, and we could say that we have a satisfactory solution to our problem. In fact, we can tidy things up a little further if we remember that

$$\exp x = e^x$$

So $\exp\left(\dfrac{t^2}{2} + d\right) = e^{(t^2/2)+d} = e^{t^2/2} \times e^d$

If we write $A = e^d$, the general solution becomes

$$q = Ae^{t^2/2} \text{ or } A \exp\left(\frac{t^2}{2}\right),$$

where $A \in R^+$, since $\exp d$ cannot be negative or zero.

The family of solution curves is illustrated below.

Graphs of $Q(t)=A\exp\left(\frac{t^2}{2}\right)$ for $A=2,1$ and 0.3
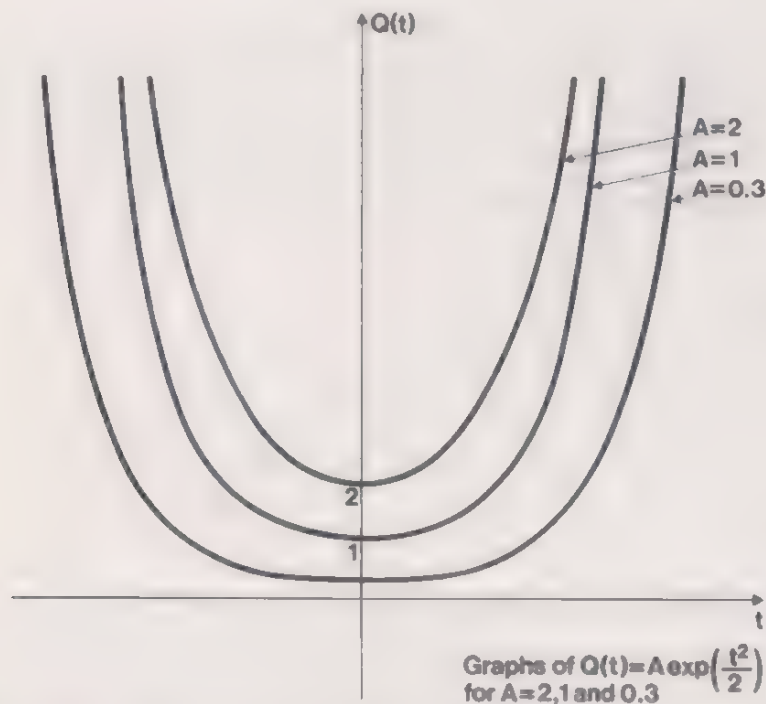
We have solved this equation in small steps, and it is worth noticing that we have sprinkled conditions on the way. Originally we required $Q(t) \neq 0$, and then we required $Q(t) > 0$, which superseded the first condition. Finally, we have

$$Q(t) = A \exp\left(\frac{t^2}{2}\right)$$

and since $A \in R^+$, the condition $Q(t) > 0$ is automatically satisfied. Also notice how the constants of integration, c and c_1, which originally appeared

added on to the ends of our expressions and were unrestricted, became "entangled" as a result of our manipulations, so that our final constant, A, is qualified ($A > 0$). With practice, the formal manipulation can become deceptively easy, but the most adept manipulators often forget to consider the consequences of what they are doing.

And yet, having been so ponderously careful, what if we had not been so careful? You can easily check that

$$Q(t) = A \exp\left(\frac{t^2}{2}\right)$$

with $A \in R^-$, also belongs to the solution set of the original differential equation. (The solution curves are similar to those described in the last figure but they are all upsidedown below the t-axis.) So perhaps a little carelessness pays. Well, it may do, just as in any other scientific field; but it must be *disciplined* carelessness. In this case, the discipline lies in checking that our formal manipulations, without conditions, have not caused us to include in our solution set functions which are not in fact solutions.

To generalize the approach using the separation of variables, we need to be able to recast the original differential equation in the form

$$(g \circ Q) \times DQ = h, \qquad\qquad\qquad \text{Equation (1)}$$

or, in terms of variables (images):

$$g(q)\frac{dq}{dt} = h(t),$$

and then we hope to be able to integrate. Herein lies the reason for the name of the method; we "separate" the q's to one side of the equation to enable us to integrate. This means that the original differential equation

$$\frac{dq}{dt} = f(t, q)$$

must be expressible in the form

$$\frac{dq}{dt} = \frac{h(t)}{g(q)}$$

Exercise 1

There are some fairly sophisticated conditions on g and h for this method to work, which we shall not discuss in this text; but what *obvious* restriction must we place on g?

Returning to the form in Equation (1), we have

$$(g \circ Q) \times DQ = h$$

Using the result for integrating a composite function quoted above, we get

$$\int (g \circ Q) \times DQ = \int h + (t \longmapsto c)$$

that is

$$\left(\int g \right) \circ Q = \int h + (t \longmapsto c)$$

Thus, if we can obtain $\int g$ and $\int h$ from appropriate tables (perhaps after further manipulation), we can solve the differential equation.

Summary

If

$$DQ = \frac{h}{g \circ Q} \qquad (g(Q(t)) \neq 0, t \in R),$$

then the solutions satisfy the equation

$$\left(\int g \right) \circ Q = \int h + (t \longmapsto c) \qquad c \in R.$$

The Leibniz form of the separation of variables rule is outlined below.

If

$$\frac{dq}{dt} = \frac{h(t)}{g(q)} \qquad (g(q) \neq 0),$$

then the solutions satisfy the equation

$$\int g(q)\, dq = \int h(t)\, dt + c$$

This is a form you will probably find in textbooks; you may find this form easier to remember.

Exercise 2

Find the function Q which is a solution of

$$q\frac{dq}{dt} = 2t \qquad (t \in R)$$

where $q = Q(t)$, and which satisfies the initial condition

$$Q(0) = 3$$

6.5   Formula Method 2: Integrating Factor

We can now find formula solutions for first order differential equations of two types

(i) $Q'(t) = g(t)$,

by straightforward integration, and

(ii) $Q'(t) = \dfrac{g(t)}{h(Q(t))} \qquad h(Q(t)) \neq 0,$

by separating the variables, *provided* the resulting integrals can be found with reasonable ease. Obviously there will be cases where the equation does not admit of the particular form required in the separation of variables method. For example,

$$Q'(t) = -Q(t) + t^2 \exp(-t)$$

cannot be put in the required form. We can separate the variables in the sense that we can get all the dependent variables on one side, i.e.

$$Q'(t) + Q(t) = t^2 \exp(-t),$$

but we cannot get the left-hand side into the form $(g \circ Q) \times DQ$. However, we can adopt the same idea. Our separation of variables method relied on our being able to use known results about the integration of composite functions, and these were themselves obtained from our result for differentiating composite functions. Now we also know how to differentiate the sum and product of functions. The "sum" result is so obvious that it is unlikely to give us any technique for solving differential equations which is not itself obvious. But the "product" result

$$D(f \times g) = (Df) \times g + f \times (Dg)$$

is not so obvious. Suppose that we assume that the left-hand side of our

differential equation is of the form $D(f \times g)$, for some choice of f and g. Then

$$(Df) \times g + f \times (Dg) = Q' + Q$$

Now we can choose $f = Q$; then we have

$$Q' \times g + Q \times (Dg) = Q' + Q$$

This equation is certainly satisfied if we can choose g so that

$$Dg = t \longmapsto 1 \qquad (t \in R)$$

and

$$g = t \longmapsto 1 \qquad (t \in R)$$

but this is plainly impossible. In general, this is in fact the only choice that we can make for g. To explain this point properly we need the idea of linear independence (which we introduce in Volume III). For the present, it is sufficient to note that we cannot, in general, make the choice that we want for g: so we must continue our search for a method.

However, now that we have started on this way of thinking, let's continue with it. We obtained the form

$$Q'(t) + Q(t) = t^2 \exp(-t)$$

by separating out the terms involving the dependent variable $q = Q(t)$; and then we tried to see if the left-hand side had product form. The latter part is the important bit, so let's concentrate on that. Can we find a left-hand side, variables separated or not, such that we can recognize a product form? Assuming that one function in the product is Q, say f as above, then we require a form

$$Q' \times g + Q \times Dg.$$

Comparing this with what we have actually got in the differential equation, we must somehow find a way of multiplying Q' by a function g, and Q by its derivative Dg. Now we can't multiply Q' by one function and Q by another without changing the differential equation: we must multiply them both (and of course the right-hand side) by the *same* function. This function must therefore satisfy the equation

$$Dg = g$$

Now any function g satisfying this equation will do, and we know one solution, the exponential function, since

$$D(t \longmapsto \exp t) = t \longmapsto \exp t$$

So we multiply both sides of our equation by $\exp t$ (which luckily is not zero anywhere, so we are not likely to produce nonsense). We obtain

$$(\exp t)Q'(t) + (\exp t)Q(t) = t^2(\exp t)(\exp(-t)) = t^2$$

i.e.

$$D(\exp \times Q) = t \longmapsto t^2,$$

or

$$\exp \times Q = t \longmapsto \frac{t^3}{3} + c,$$

from which we can get Q or $Q(t) = q$ explicitly as

$$q = \exp(-t) \times \left(\frac{t^3}{3} + c \right)$$

So the solution set is

$$\left\{ Q : Q(t) = \exp(-t) \times \left(\frac{t^3}{3} + c \right) \quad (t \in R), c \in R \right\}.$$

Before we discuss some general aspects of this method we suggest you try it for yourself in the following exercise.

Exercise 1

Find the solution set of the equation

$$\frac{dq}{dt} + \frac{q}{t + a} = t \qquad (t \in R, t \neq -a)$$

where a is a positive number and $q = Q(t)$.

The crucial step in both the example in the text and Exercise 1 is the multiplication by a suitable function g, so that the left-hand side of the equation can be written as the derivative of the product of g with Q, i.e.

$$D(g \times Q) = gQ' + g'Q$$

If we are going to adopt this as the general form, then it implies that we are dealing with an equation of the form

$$Q' + \frac{g'}{g}Q = \frac{h}{g}, \qquad g(t) \neq 0,$$

except that we are not going to be told what g is.

So let's start with a given equation of the general form

$$Q' + PQ = R,$$

where P and R are known functions. Comparing this with the previous equation, we see that g is determined by

$$\frac{g'}{g} = P$$

Remembering that we are looking for just one solution, we ignore the constant of integration, and get

$$\ln \circ g = \int P \qquad g(t) \in R^+$$

or

$$g = \exp \circ \int P$$

or

$$g(t) = \exp\left(\int P(t)\,dt\right)$$

Summarizing, we get the following rule:

To solve an equation of the form

$$Q'(t) + P(t)Q(t) = R(t),$$

where P and R are known functions, multiply throughout by the *integrating factor*

$$g(t) = \exp\left(\int P(t)\,dt\right)$$

and integrate directly.

Let us try an example. We shall find the solution set of the equation

$$Q'(t) + 2tQ(t) = 4t^3 \exp(-t^2) \qquad (t \in R)$$

To compare with the rule, we take

$$P : t \longmapsto 2t$$

and a simple integral is

$$t \longmapsto t^2.$$

Therefore an integrating factor is

$$g(t) = \exp\left(\int P(t)\,dt\right) = \exp(t^2) = e^{t^2},$$

and we have

$$\frac{d}{dt}(e^{t^2}Q(t)) = e^{t^2}Q'(t) + 2te^{t^2}Q(t) \qquad \text{(derivative of a product)}$$

$$= e^{t^2}(Q'(t) + 2tQ(t))$$

$$= e^{t^2}(4t^3e^{-t^2}) \qquad \text{(from given differential equation)}$$

$$= 4t^3$$

(Note that the left-hand side is $\dfrac{d}{dt}$ (integrating factor $\times Q(t)$).)

The integral is then

$$Q(t)e^{t^2} = t^4 + c,$$

and the solution set may be written as

$$\{Q : Q(t) = e^{-t^2}(t^4 + c) \qquad (t \in R)\, c \in R\}$$

Exercise 2

Indicate which of the following differential equations can be solved (assuming that the appropriate primitive functions can be found for the integrals involved), by

A: both the method of separation of variables and the integrating factor method;

B: the method of separation of variables and not the integrating factor method;

C: the integrating factor method and not the method of separation of variables;

D: neither the integrating factor method nor the method of separation of variables.

(i) $\dfrac{dq}{dt} = \ln(tq) \qquad (t \in R^+, q = Q(t) \in R^+)$

(ii) $(\exp t)\dfrac{dq}{dt} = q^2$

(iii) $Q'(t) = \sqrt{t}Q(t) \qquad (t \in R^+, Q(t) \in R^+)$

(iv) $\dfrac{dq}{dt} = \cos q + \cos t$

(v) $\dfrac{dq}{dt} = \dfrac{q^2 + 1}{t^2 + 1}$

Solve two of the above equations for which you can find the primitive functions for the integrals involved.

6.6 Additional Exercises

Exercise 1

Find the solution set of

$$\frac{dq}{dt} = \frac{-kq}{q + a}.$$

where $q = Q(t)$, and a and k are positive numbers. State any conditions you find necessary in the manipulations.

Exercise 2

Find the solution set of

$$t\frac{dq}{dt} + 2(q - 4t^2) = 0 \qquad (q, t \in R)$$

where $q = Q(t)$.

6.7 Answers to Exercises

Section 6.2

Exercise 1

For example,
(i) $t^2 - 3t + 2 = 0$ whose solution set is $\{2, 1\}$.
(ii) $t^2 - 2t + 1 = 0$ whose solution set is $\{1\}$.
(iii) $t^2 - 2t + 2 = 0$ whose solution set has no real members.

Exercise 2

The solution set can be written in the form

$$\{n\pi : n \in Z\},$$

since the sine of any integer multiple of π is zero. There are as many solutions as there are positive integers. The solutions can be put in one–one correspondence with the set of all positive integers as follows:

$$0 \longleftrightarrow 1$$

$$\pi \longleftrightarrow 2$$

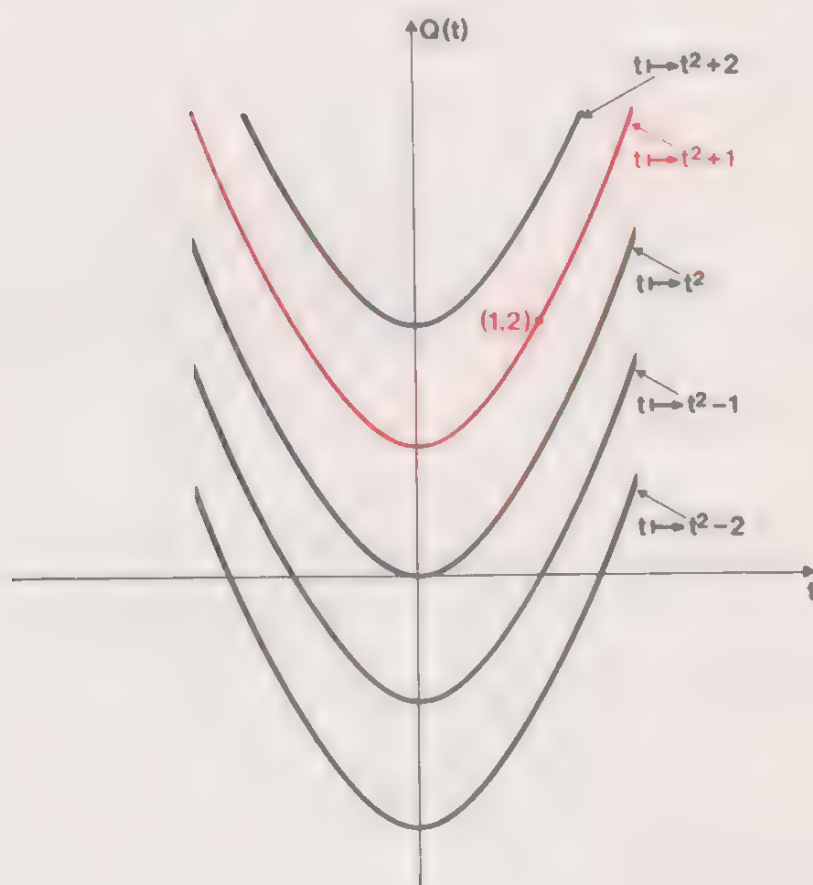$$-\pi \longleftrightarrow 3$$

$$2\pi \longleftrightarrow 4$$

$$-2\pi \longleftrightarrow 5$$

and generally,

$$k\pi \longleftrightarrow 2k \qquad \text{for } k > 0,$$

$$k\pi \longleftrightarrow -2k + 1 \quad \text{for } k < 0.$$

Exercise 3

The solution set is

$$\{Q : Q(t) = t^2 + c \qquad (t \in R), c \in R\}.$$

Using the initial condition

$$Q(1) = 2 = 1 + c$$

we obtain

$$c = 1$$

and the particular solution curve corresponds to

$$Q(t) = t^2 + 1.$$

The family of curves is, in this case, a family of parabolas. All the curves look similar in the same way that a biological family may have similar traits. That is one reason why it is convenient to use the word *family* to describe this particular collection of curves.

Exercise 4

(i) $\qquad Q' = t \longmapsto 2t \qquad (t \in R)$

$\qquad$ or $\quad DQ = t \longmapsto 2t \qquad (t \in R)$

(ii) $\qquad \dfrac{dq}{dt} = 2t$

Section 6.4

Exercise 1

$g(q)$ may not be zero, i.e. we must exclude those values of t for which $g(Q(t)) = 0$.

Exercise 2

The equation is already in separated form. Comparing it with the equation

$$\frac{dq}{dt} = \frac{h(t)}{g(q)}$$

we see that

$$g(q) = q, h(t) = 2t$$

Therefore suitable primitive functions of g and h respectively are

$$\int q \longmapsto q = q \longmapsto \frac{q^2}{2} \quad \text{and} \quad \int t \longmapsto 2t = t \longmapsto t^2$$

So the solution set is

$$\left\{ Q : \frac{q^2}{2} = t^2 + c \qquad (t \in R), c \in R \right\}.$$

The particular value of c determined by the initial condition is given by

$$\tfrac{1}{2}(Q(0))^2 = 0^2 + c, \text{ where } Q(0) = 3,$$

i.e.

$$c = \tfrac{9}{2}$$

Therefore the particular solution we are interested in is given by

$$\frac{(Q(t))^2}{2} = t^2 + \frac{9}{2}$$

But this determines two possible functions Q:

(i) $Q : t \longmapsto \sqrt{2t^2 + 9}$

(ii) $Q : t \longmapsto - \sqrt{2t^2 + 9}$

Remembering again that $Q(0) = 3$, i.e. $Q(0) > 0$, we see that the required solution is

$$Q : t \longmapsto \sqrt{2t^2 + 9}$$

Section 6.5

Exercise 1

As before, we assume that we can get the left-hand side of the equation in the form of a derivative of a product, i.e. we assume that it can take the form

$$\frac{d}{dt}(q \times g(t)) = \frac{dq}{dt} g(t) + q \frac{d(g(t))}{dt}$$

To get the left-hand side of our differential equation into this form, we must multiply $\frac{dq}{dt}$ by $g(t)$ and $\frac{q}{t + a}$ by $\frac{d(g(t))}{dt} \times (t + a)$. These two must be the

same, i.e. we need a function g such that

$$(t + a)\frac{d(g(t))}{dt} = g(t)$$

We can separate the variables in this equation in the sense of our first method, to obtain

$$\frac{1}{g(t)}\frac{d(g(t))}{dt} = \frac{1}{t + a}$$

We are looking for any one solution of this equation so we take the constant of integration to be zero.
(We have seen the left-hand side before.)
Integrating, we get one solution

$$\ln (g(t)) = \ln (t + a) \qquad (t + a > 0)$$

or

$$g(t) = t + a$$

So multiplying both sides of the original differential equation by $t + a$ (which again is not zero, since $t + a > 0$), we get

$$(t + a)\frac{dq}{dt} + q = t(t + a)$$

(Of course, this rearrangement may have been obvious to you from the start: we have given a formal approach just in case it wasn't.) We have

$$\frac{d}{dt}(q(t + a)) = t^2 + at,$$

so that

$$q(t + a) = \frac{t^3}{3} + \frac{at^2}{2} + c,$$

which gives the solution set

$$\left\{ Q : Q(t) = \frac{1}{t + a}\left(\frac{t^3}{3} + \frac{at^2}{2} + c\right)(t \in R, t \neq -a), c \in R \right\}.$$

(In the process of solution we required $t + a > 0$, but this condition is not essential: it is easy to verify that the above describes the solution set for $t + a < 0$ as well.)

Exercise 2

(i) *D*

(ii) *B*

Solution set is

$$\left\{ Q : Q(t) = \frac{1}{\exp(-t) + c} \qquad (t \in R), c \in R \right\}$$

If c is negative, there will be the problem that $Q(t)$ is not defined for one value of t and so there would be a discontinuity in Q.

(iii) *A*

Solution set is

$$\{ Q : Q(t) = \exp(\tfrac{2}{3}t^{3/2} + c) \qquad (t \in R^+), c \in R \}$$

or

$$\{ Q : Q(t) = a \exp(\tfrac{2}{3}t^{3/2}) \qquad (t \in R^+), a \in R^+ \}$$

(iv) *D*

(v) *B*

To find the solution by the method of separation of variables requires us to be able to recognize the primitive function

$$\int t \longmapsto \frac{1}{t^2 + 1}$$

We have not formally covered this in the text: it is, in fact,

$$t \longmapsto \text{arc tan } t \qquad (t \in R)$$

which is defined to be the inverse function of

$$t \longmapsto \tan t \qquad \left(-\frac{\pi}{2} < t < \frac{\pi}{2} \right).$$

The solution set is then

$$\{ Q : \text{arc tan } Q(t) = \text{arc tan } t + c \qquad (t \in R), c \in R \}.$$

This could be simplified if there were any further interest in the problem.

Section 6.6

Exercise 1

A first condition is obviously that $q + a \neq 0$, i.e. $Q(t) \neq -a$. In separated form the equation becomes

$$\frac{q + a}{q} \frac{dq}{dt} = -k,$$

and we have to impose the further condition $q \neq 0$, i.e. $Q(t) \neq 0$. Referring back to the summary of section 6.4, we have

$$g(q) = \frac{q + a}{q} = 1 + \frac{a}{q}, \, h(t) = -k$$

Suitable primitive functions are

$$\int g = q \longmapsto q + a \ln q \qquad \text{(assuming } q \in R^+\text{)},$$

$$\int h = t \longmapsto -kt$$

Therefore, the solution set is

$$\{Q : Q(t) + a \ln Q(t) = -kt + c \qquad (t \in R), c \in R, Q(t) \in R^+\}$$

Notice that $Q(t) \in R^+$ takes care of the previous restrictions on $Q(t)$. This time we cannot get the solution set in an explicit form, that is, we cannot get q, the dependent variable, explicitly in terms of t, the independent variable, and thus it is difficult to determine $Q(t)$ for a given t.

Exercise 2

The equation may be rearranged as

$$Q'(t) + \frac{2}{t} Q(t) = 8t.$$

Integrating factor is

$$\exp\left(\int \frac{2}{t} dt\right) = \exp(2 \ln t) = t^2$$

Multiplying through by the integrating factor gives

$$t^2 . Q'(t) + 2t . Q(t) = 8t^3$$

which has solution

$$t^2 . Q(t) = 2t^4 + c.$$

The solution set may thus be written

$$\{Q : Q(t) = 2t^2 + ct^{-2} \qquad (t, c \in R)\}.$$

CHAPTER 7 APPROXIMATION

7.0 Introduction

In the calculus we are essentially concerned with the mathematics of the set of real numbers (R). The techniques of calculus can be used in the analysis of problems but will serve only to provide solutions to the problems in terms of formal mathematical expressions. If we require a numerical solution we are obliged to carry out numerical computation.

For example, $\int x \longmapsto \sin x$ can be expressed formally as $x \longmapsto \cos x + c$

but $\int_0^1 x \longmapsto \sin x$ requires evaluation since it is a number, and the evaluation from $x = 0$ to $x = 1$, at intervals of 0.1, will produce a table of values.

We are obliged in computation to replace real numbers by rational numbers, which have a finite representation. In this chapter we consider the problems which arise when we round numbers in this way and therefore cause errors in our calculated answers.

We also look at other sources of error, such as errors in the initial data (we exclude actual blunders) and notice how these and rounding errors develop after even the simplest operations. For although the sources of error are essentially static, their growth during a computation is a dynamic process and estimation of the resulting propagation of the errors is a major and interesting problem. We have occasionally remarked on such errors and used some of the associated ideas extensively, but in this chapter we look at them explicitly.

The chapter continues with a look at some of the techniques used to supply values missing from a table of data, which may consist either of observed values or of discrete values of a function. from which we wish to obtain an approximate continuous function.

You will note that we have avoided using the techniques of calculus in our discussion of the problems and have in general preferred to treat the questions from first principles. There are however obvious links between the material of this chapter and the work on approximation in Chapter 5.

7.1 Types of Error

We begin by considering the problem of getting accurate answers in numerical computation and by examining the ways in which errors can arise and with the effect of these errors upon the accuracy of the calculated answers.

"Can you guess the weight of the cake?" "How many dried peas in the jar?" Have you ever been asked these questions at a local fête? How accurately can you weigh a cake with your hand? To the nearest kilogram? Can you estimate the number of dried peas to the nearest 100? Embodied in these simple examples is the idea that many of the numbers that occur in our life are approximations. At what time did you leave the house this morning? How many cars did you see on your way to work? It is doubtful whether you could answer either of these questions precisely. The law recognizes that inaccuracies are inevitable. If you obtain 5 gallons of petrol from a petrol pump which has been in use for some time, legally the quantity you get may be between $4\frac{63}{64}$ and $5\frac{1}{16}$ gallons. Even the extremely precise atomic clocks have a possible inaccuracy of 5 seconds in 700 years.

The various quantities quoted above fall into two types. One type comprises the quantities, like the weight of the cake, that we can never find precisely; no matter how fine a balance we use there is always a small possible error in the measurement of the weight. The other type comprises the quantities, like the number of peas in the jar, that we can, in principle, find precisely by counting. In this text we are concerned mainly with quantities of the first type. All these are examples of measurement errors arising because some measurement of a physical quantity is not perfectly accurate. Measurements are not, however, the only source of inaccuracies: try writing π or $\frac{1}{3}$ as an exact decimal. However many decimal places you write down there must be some error in your representation. Such an error, arising from the fact that the number is not given exactly by the decimal representation used, is called a round-off error. Errors of measurement and round-off errors in data have a similar effect when the data are used in a calculation. However, it should be stressed that measurement errors normally arise from many individual causes which combine together in a random way, so that one cannot state the absolute limit of a measurement error. All that one can do is to say how likely it is that a given error has not been exceeded. We refer to the two types of error collectively as *inherent errors*.

Let us take the above example of the dried peas in the jar a little further. Suppose you are not allowed to count them but would like to improve

on a simple guess. So you argue something like this: on average, a dried pea looks as if it is $\frac{1}{2}$ cm in diameter. The jar has a square base with sides about 8 cm long, and about 10 cm high. Thus you deduce that the number is somewhere in the region of

$$\frac{8}{\frac{1}{2}} \times \frac{8}{\frac{1}{2}} \times \frac{10}{\frac{1}{2}} = 5120$$

Let us analyse briefly what you have done. You have used inaccurate data in an exact computation and derived, as you know, an inaccurate result. The question arises: "How inaccurate is the result?" and this takes us into the topic of the propagation of errors. By this we mean the way in which errors in the initial data used in a computation affect the final result and any intermediate results.

This topic is introduced in this chapter by investigating firstly the propagation of errors by functions with domain and codomain R.

Exercise 1

How many dried peas would you estimate to be in the jar in the previous example if you assumed the diameter of the pea to be 0.4 cm? Are you surprised at the different result you obtain?

7.2   Absolute and Relative Error

It may seem surprising that a mathematical treatment of errors can exist. One naturally thinks of mathematics as an exact discipline, in which errors can arise only through mistakes or imperfections which should not be tolerated in mathematical work. This is a misconception, however; provided we can define precisely what we mean by an "error" and attach a numerical value to it, we can apply mathematical reasoning to the errors just as we do with any of the other objects to which we apply mathematics.

To define "error" mathematically, let us suppose that we are using one number, which we denote by x, as an approximation to another, which we denote by X. For example, the "exact" number X might be $\frac{1}{3}$ and x the approximation 0.33; or X could be π and x the approximation $\frac{22}{7}$; or X could be the actual number of peas in a jar and x your guess at this number; or X could be the precise amount of petrol you received and x the amount as measured by the meter on the petrol pumps. In each case, the numbers X and x are likely to be different, and we define their difference $x - X$ as the absolute error in x and denote it by e_x:

$$e_x = x - X$$

The word "absolute" is to distinguish this measure of error from another one, called the "relative error", which we shall meet presently. (In general, we use just "error" when it is clear from the context which we mean.)

Note that e_x can be either positive or negative, according as the approximation x is larger or smaller than the exact value X. As an example, if we imagine we have a worn tape measure with the first centimetre missing, then if the true length being measured were 14 cm, we would, assuming for a moment that the rest of the tape measured exactly, record a value of 15 cm. This we would call the approximate value in this instance. Thus

$$15 \text{ cm} \quad - \quad 14 \text{ cm} \quad = \quad 1 \text{ cm}$$

(approximate value − true value = absolute error)

and the absolute error would be 1 cm.

In other words, to correct the values given by the tape, we must always make a correction (the negative of the error) of −1 cm, i.e. we *subtract* the error.

For another example, if you know that your speedometer consistently records 5 mile h too high, a recorded (approximate) value of 37 mile/h would correspond to a true value of 32 mile h with an absolute error of 5 mile h and a correction needed of −5 mile h, i.e.

$$37 \text{ mile/h} \quad - \quad 32 \text{ mile/h} \quad = \quad 5 \text{ mile/h}$$

(approximate value − true value = absolute error)

In the example of the tape measure above, we had an error of 1 cm in 15 cm. It is possible to measure a distance of 1 km to an accuracy of 1 cm, i.e. the possible absolute error is again 1 cm in a recorded value of 1 km. Clearly this second measurement is, in a sense, more accurate than the first, although the absolute error is the same in each case. To allow for this type of distinction we use the relative error in x defined as $\frac{e_x}{x}$ and written as r_x:

$$r_x = \frac{e_x}{x}$$

(Note that we compare e_x with x, the approximate value, because in general we know this value and not the exact value X.)

Thus in the above two examples we have approximate value $x = 15$ cm,

absolute error $e_x = 1$ cm, giving

$$\text{relative error } r_x = \frac{1}{15}$$

and approximate value $x = 10^5$ cm, absolute error $e_x = 1$ cm, giving

$$\text{relative error } r_x = \frac{1}{10^5} = 10^{-5}$$

showing how much smaller the relative error is in the second case. Multiplied by 100, the relative error is the percentage error you have probably met before. Often knowledge of the relative, or percentage, error is more useful than knowledge of the absolute error, since it gives a measure of the error in relation to the size of the number being considered. This is not always the case however; for example, the absolute error in the diameter of an axle is clearly the more important when we are fitting it into a ball-race. And if the approximate value of the number is zero (as when measuring the oxygen content of polluted river water!), the definition of relative error loses its meaning.

Accuracy

The naive answer to the question: "What is accuracy?" is that it is simply the absence of error — i.e. that a small error corresponds to a high accuracy and vice-versa. There is a lot of truth in this, but it is not the whole story. Suppose you bought a nominal 5 gallons at each of two apparently identical petrol pumps designed to comply with the legal requirement mentioned earlier, i.e. that the true amount must lie between $4\frac{63}{64}$ and $5\frac{1}{16}$ gallons, and that at one pump you happened by chance to get $5\frac{1}{160}$ gallons, and at the other you got the legal maximum, $5\frac{1}{16}$ gallons. At the first pump the error was $\frac{1}{160}$ gallon and at the second it was $\frac{1}{16}$ gallon, but would it be reasonable to say that one pump was ten times as accurate as the other on this account alone? On another day, the position might be reversed, purely by chance. We would like to frame our definition of accuracy so as to be independent of such caprices.

We can do this by making the definition of accuracy depend on the pump itself and not on the amount it delivers on any particular occasion. That is, the accuracy is defined by specifying, not the error on any par-

ticular occasion, but bounds between which the error must lie. In the case of the petrol pump, the accuracy in measuring 5 gallons must satisfy the legal requirement that x, the amount of petrol delivered, must lie between $4\frac{63}{64}$ and $5\frac{1}{16}$. By our definition of the absolute error, $e_x = x - X$, this condition requires that e_x lie between $-\frac{1}{64}$ and $+\frac{1}{16}$. In symbols, this is

$$-\tfrac{1}{64} \leqslant e_x \leqslant \tfrac{1}{16}$$

It can also be written

$$e_x \in \left[-\tfrac{1}{64}, \tfrac{1}{16}\right]$$

where $\left[-\frac{1}{64}, \frac{1}{16}\right]$ is the set consisting of all numbers from $-\frac{1}{64}$ to $\frac{1}{16}$ inclusive. Such a set is called an interval, and in this particular context it is called an error interval.

This provides the answer to our question: "What is accuracy?" The accuracy of an approximate number is specified by giving an interval within which the error in the number must lie. The reason why we specify the accuracy in this way, rather than by giving the error itself, is that we do not normally know the error — if we did, we could just subtract it from the approximate number and recover the exact number.

It frequently happens (though not in the petrol pump example) that the interval used to specify the accuracy is symmetrical about zero so that the condition on the error in a number x has the form

$$e_x \in [-\varepsilon_x, \varepsilon_x]$$

where ε_x is some positive number.
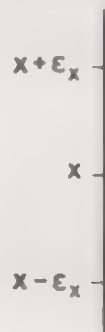
This condition can also be written

$$|e_x| \leqslant \varepsilon_x$$

When the accuracy is specified by a symmetrical interval like this, we call the number ε_x the absolute error bound of x. An alternative way of specifying the accuracy of an approximate number x is to use the relative error bound defined by

$$\rho_x = \frac{\varepsilon_x}{|x|}$$

so that the relative error r_x satisfies

$$|r_x| \leqslant \rho_x$$

Exercise 1

We record a measurement of 2.5 kg and assume that there is a maximum error in the instrument of 0.05 kg, that is, the true value is in the interval [2.45, 2.55] kg.

What is

(i) the absolute error bound?
(ii) the relative error bound?

A common notation for specifying absolute error bounds is to write, for example,

$$\pi = 3.14 \pm 0.005$$

to indicate that 3.14 is an approximate value for the exact number π and that the absolute error bound is 0.005.

Another method of specifying error bounds depends on the convention for rounding off decimals. You probably know already how to round off a decimal to fewer places. For example, the number π to 10 decimal places is

3.1415926536 ...

To save writing and arithmetical labour we very rarely work with this value, but use the best approximation obtainable with, say, 2 or 4 decimal places. The two-place approximation is

3.14

since this has an error

$$3.14 - 3.1415926536\ldots = -0.0015926536\ldots$$

whereas any other two-place approximation, say 3.13 or 3.15, would have a larger error. In this case the two-place approximation is identical with the first 3 digits of the exact (non-terminating) decimal for π. With four places, on the other hand, the best approximation is

$$3.1416$$

since the error

$$3.1416 - 3.1415926536\ldots = 0.0000073464\ldots$$

is smaller than for (say) 3.1415 or 3.1417. This procedure of representing a number by the closest decimal with some given number, say n, of digits after the decimal point, is called rounding-off the number to n decimal places.

If an exact number, X, is approximated by its round-off form with n decimal places, x_n, the absolute error bound is

$$\overbrace{0.00\ldots05}^{n \text{ zeros}}$$

since

$$X \in [x_n - \overbrace{0.00\ldots05}^{n \text{ zeros}}, \; x_n + \overbrace{0.00\ldots05}^{n \text{ zeros}}]$$

This shows that any rounded-off decimal implies an error bound, and so we can use rounded-off decimals to specify the accuracy of an approximation without giving the error bound explicitly. Thus we write

$$\pi = 3.14$$

or

$$\pi = 3.14 \text{ to two decimal places}$$

to mean that the approximation 3.14 has the error bound characterizing two-place accuracy, i.e. that

$$\pi = 3.14 \pm 0.005$$

There is one convention that should be mentioned here. When rounding a number to one less figure we increase the previous digit by 1 if the last digit is 6, 7, 8 or 9. If the last digit is 0, 1, 2, 3 or 4, we leave the previous digit untouched. If the last digit is a 5, the convention is that we look at

the previous digit: if it is *even* we leave it unchanged, if it is *odd* we increase it by 1. This is to avert any bias in always rounding to the larger number.

Exercise 2

The scale of a digital voltmeter is calibrated to an accuracy of two decimal places. Assuming that it is perfectly accurate, what are the absolute error bounds in readings of 10v, 0.1v, 0.01v?

Sometimes the term significant figures is used instead of "the number of decimal places". For example, we say that the number

12.04

has four significant figures — two in front of the decimal point and two after. The number

0.0001204

also has four significant figures, the last four. (The first three zeros only serve to distinguish the number from 0.1204, for example, and are not said to be significant.) In the statement

"The sun is 93 000 000 miles from the earth"

only the first two figures are significant: the statement means that the distance of the sun from the earth is closer to 93 000 000 than to 94 000 000 or 92 000 000, not that it is closer to 93 000 000 than to 93 000 001 or 92 999 999. To avoid ambiguity in these cases it is convenient to write such numbers in the form

$$9.3 \times 10^7$$

which makes it clear that there are just 2 significant figures, so that the absolute error bound is 0.05×10^7 or 500 000.

Exercise 3

Determine the absolute error bound, ε_x and the relative error bound ρ_x given that we record a measurement for x as follows:

(i) 700 years with a maximum error of 1 second.
(ii) 3 kilohms with a maximum percentage error of 10%.
(iii) 10 microfarads with a maximum percentage error of 20%.

Accuracy is always an important objective, but we must take care not to claim to have attained more of it than in fact we have. Very often, even where one can theoretically attain a high accuracy, it may not be worth while. For example, if your temperature were recorded as 39.962 °C rather than 40 °C the extra digits would convey no more information to you or the doctor. You would just be very sick. In any case shortly afterwards the temperature could well have changed considerably from the former value whilst remaining approximately 40 °C.

Care must always be taken in any calculation (both to ensure the credibility of the result and to save work) to quote the result only to the accuracy implied by the data and the calculation process. You may not be able to do this precisely now, but at least you should be able to recognize when the result is clearly overstated. This is frequently referred to in the sciences as recognizing the *order of magnitude* of the errors involved. An instance of striving for accuracy which is unattainable is illustrated by the following exercise.

Exercise 4

In the right angled triangle shown we measure the height as 1 metre and the base as 2 metres, both measurements being accurate to the nearest centimetre. By the theorem of Pythagoras the length of the hypotenuse can be calculated as 2.23607 metres. Is this a sensible deduction?
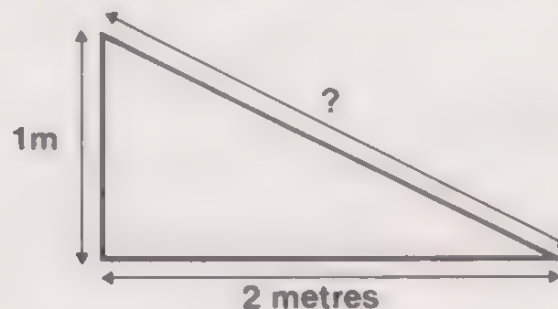
7.3 Propagation of Errors

In mathematics we are concerned, not only with numerical data, but also with calculations that may be performed on the numbers forming the data. If there are errors in the data, they will affect the result of the calculation, and so the accuracy of the result depends on the accuracy of the data. The mathematical theory of errors makes it possible to express the accuracy of the result of a given calculation in terms of the accuracy in the data. In this section you will learn how to do this in the

simple case where the calculation in question is the evaluation of images of real functions such as

$$f : x \longmapsto 3x^2 - 2x + 1 \qquad (x \in R)$$

or

$$g : x \longmapsto 5x^4 - 6x + \frac{1}{x} - \frac{7}{x^3} \quad (x \in R, x \neq 0)$$

or

$$h : x \longmapsto \frac{1}{1 + x} \qquad (x \in R^+)$$

in which the formula specifying the rule contains only integer powers of x in the numerator and/or the denominator. We will call the set of such functions P.

In section 7.2 we defined absolute error e_x and relative error r_x in terms of the true value X and the approximate value x. These errors can occur in numbers which we may have to use in subsequent calculations, for example, when we evaluate the images of such numbers under a function from P. The question we wish to answer is: What happens to these errors when we evaluate these images? Given an error e_x in the original number, what is the error in the image? Suppose for the moment that we know the true value X of the original number and that the approximate value is

$$x = X + e_x$$

Diagrammatically we have

$$x(= X + e_x) \longrightarrow \boxed{f} \longrightarrow f(x)(= f(X) + ?)$$

The exact value of the image is $f(X)$, and the approximate value we obtain if we use x instead is $f(x)$, so that the error in the image is

$$e_{f(x)} = f(x) - f(X) \hspace{3cm} \textbf{Equation (1)}$$

It should be pointed out, however, that we may not be able to carry out an exact computation of the image even if the data itself is exact; and even if we can, it may in fact be very tedious and clumsy to have to calculate $e_{f(x)}$ in many cases. Very often the errors are small compared with x (i.e. the relative error is very small), so that if, for instance, the square of the error occurs in $e_{f(x)}$ it will be smaller still. We shall see in the

following examples that we can often usefully simplify the formula for the error in the image and obtain a satisfactory estimate much more quickly than by using Equation (1).

Multiplication by an Exact Number

If

$$f : x \longmapsto 5x \quad (x \in R)$$

then we can represent the action of a function f by the diagram

$$x = (X + e_x) \longrightarrow \boxed{x \longmapsto 5x} \longrightarrow 5X + 5e_x$$

and represent the corresponding absolute errors in the domain and codomain schematically by

$$e_x \longrightarrow 5e_x = e_{5x}$$

> To find the absolute error in the image we simply multiply the absolute error in the original number by the appropriate factor.
>
> **Rule 1**

Other Products

Consider the function "square it".

$$f : x \longmapsto x^2 \quad (x \in R)$$

Then

$$x = (X + e_x) \longrightarrow \boxed{x \longmapsto x^2} \longrightarrow (X^2 + 2Xe_x + e_x^2)$$

and

$$e_x \longrightarrow 2Xe_x + e_x^2 = 2xe_x - e_x^2 = e_{x^2}$$

The second expression on the right-hand side is found by substituting

$$X = x - e_x$$

in the second term of the first expression. Again, with the particular numerical values $X = 1, e_x = 0.1$, we find

$$(1 + 0.1) \longrightarrow \boxed{x \longrightarrow x^2} \longrightarrow 1 + 2 \times 1.1 \times 0.1 - 0.01$$

$$= 1 + 0.22 - 0.01$$

$$= 1 + 0.21$$

Note the relative sizes of the terms on the right. The number 0.01 is small compared even with the total error 0.21. This shows us where we can gain in simplicity with the marginal loss of some accuracy. Provided e_x is *small compared with* x, e_x^2 will always be *much smaller* than $2xe_x$ and we can safely ignore it. Thus we can usefully say that the error in x^2, e_{x^2}, is about $2xe_x$, i.e.

$$x \longrightarrow \boxed{x \longrightarrow x^2} \longrightarrow x^2$$

$$e_x \longrightarrow \text{estimated } e_{x^2} = 2xe_x$$

If we look at the behaviour of the relative error for the same function, we get the very simple rule

$$r_{x^2} \simeq \frac{2xe_x}{x^2} = 2r_x \quad (x \neq 0)$$

> Squaring an approximate number roughly doubles its relative error.
>

In the same way we can find a useful estimate of the absolute error in x^3 if the absolute error in x, e_x, is small compared to x. The value we obtain is $e_{x^3} \simeq 3x^2 e_x$.

We can then express r_{x^3} approximately in terms of r_x, assuming $x \neq 0$:

$$r_{x^3} = \frac{e_{x^3}}{x^3} \simeq \frac{3x^2 e_x}{x^3} = \frac{3e_x}{x} = 3r_x$$

Generalizing from x^2 and x^3 suggests

$$e_{x^n} \simeq nx^{n-1}e_x, \qquad r_{x^n} \simeq nr_x$$

This is an important result, which can be justified by applying the Binomial Theorem to expand $(X + e_x)^n$ and then taking e_x small compared to x so that terms in e_x^2 and above can be ignored.

From this result we can conjecture an important principle, i.e. that in multiplication we can *add* relative errors to obtain an *estimate* of the relative error in the product.

This is an important point to remember for future use.

> For an estimate of the relative error in multiplication, add the relative errors.
>
> **Rule 3**

Division

Consider

$$f : x \longmapsto \frac{1}{x} \qquad (x \in R, x \neq 0)$$

Then

$$(X + e_x) \longrightarrow \boxed{x \longmapsto \frac{1}{x}} \longrightarrow \frac{1}{X + e_x} = \frac{1}{X} + ?$$

By some algebraic manipulation (you need not derive this, just check it if you wish),

$$e_{1/x} = \frac{1}{X + e_x} - \frac{1}{X} = \frac{-e_x}{(X + e_x)X} = \frac{-e_x}{x(x - e_x)}$$

We ignore the e_x in the denominator by comparison with the x next to it, and get

$$e_{1/x} \simeq -\frac{e_x}{x^2}$$

and

> $$r_{1/x} = \frac{e_{1/x}}{1/x} \simeq -\frac{e_x}{x} = -r_x$$
>
> **Rule 4**

Example 1

By writing

$$\frac{1}{x^2} = \left(\frac{1}{x}\right)^2$$

estimate r_{1/x^2} and e_{1/x^2}.

By using Rule 3 for error estimates in multiplication we find

$$r_{1/x^2} \simeq 2r_{1/x}$$

and hence

$$r_{1/x^2} \simeq -2r_x$$

by Rule 4.

Therefore, we have

$$e_{1/x^2} = \frac{1}{x^2}(r_{1/x^2}) \simeq -\frac{2r_x}{x^2} = -\frac{2e_x}{x^3}$$

Note that if we want the absolute error estimate of a product, it is simpler to find the relative error estimate first by the simple Rule 3.

Addition and Subtraction

It is fairly clear that for these operations we simply add (or subtract) the appropriate absolute error estimates. Thus, for example, consider the function

$$f : x \longmapsto x^3 + x^2 + x \quad (x \in R)$$

The absolute error in the image is

$$e_{x^3} + e_{x^2} + e_x \simeq (3x^2 + 2x + 1)e_x$$

Two points emerge from this:

(i) The absolute error in a sum is equal to the sum of the absolute errors in its terms. **Rule 5**

(ii) For addition and subtraction, even if we wished to find the relative error, it is simpler to find the absolute error first. **Rule 6**

Thus, in the above, the estimated relative error would be

$$\frac{3x^2 + 2x + 1}{x^3 + x^2 + x} e_x = \frac{3x^2 + 2x + 1}{x^2 + x + 1} r_x$$

Exercise 1

Estimate the absolute errors in the images of x, with absolute error e_x, under the functions

(i) $\qquad x \longmapsto x^3 - 4x + 3 \quad (x \in R)$

(ii) $\qquad x \longmapsto \dfrac{5}{x} - \dfrac{3}{x^2} \qquad (x \in R, x \neq 0)$

Summary

We summarize below the main rules we have obtained in this section for the propagation of errors in evaluating images of real functions of the type we considered.

Combination of functions	
Operation	Error estimate
Addition (or Subtraction)	Add (or Subtract) *Absolute* errors **Rule 5**
Multiplication by exact number	Multiply *Absolute** error by exact number **Rule 1**
Multiplication (or Division)	Add (or Subtract) *Relative* errors **Rules 3, 4**

The rules we have just given apply to the estimated errors themselves, not to the error bounds. It is possible to formulate rules for combining estimated error bounds, but we shall not do it here because they are more complicated than the ones for the errors.

7.4 Error Intervals

In the last section we discovered methods of estimating the error in the image of a number in the domain of a function when we know the error in that number. Generally, of course, we do not have this information;

* The relative error is unchanged in this case.

we know only that the number lies in some *interval* in the domain. Can we map this *error interval* in the domain into some *error interval* in the codomain?

Consider the mapping diagram shown.
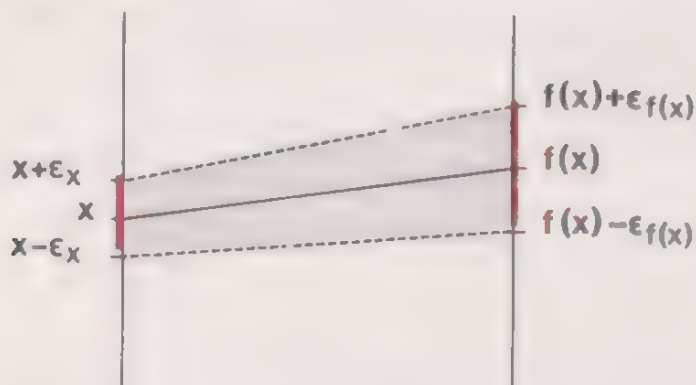
The *error interval* in the domain is known and in this particular case it is determined by the two numbers, x (approximate number) and ε_x (absolute error bound), which are known. The number x maps to the image $f(x)$ in the codomain. We can find the exact images of $x + \varepsilon_x$ and $x - \varepsilon_x$, but usually it is simpler and quicker to estimate these images by the methods of the previous section, and hence find the estimate of the error interval in the codomain from them. The dashed lines in the diagram are meant to indicate the way the *interval* maps under the function and not the images of $x + \varepsilon_x$ and $x - \varepsilon_x$. These upper and lower bounds do not necessarily map respectively to the upper and lower bounds of the image error interval as we see in the next example.

Example 1

(i) Determine the error interval in the codomain which corresponds to the error interval $[0.4, 0.6]$ in the domain under the mappings:

(a) $x \longmapsto x^2 \qquad (x \in R)$,

(b) $x \longmapsto \dfrac{1}{1 + x} \quad (x \in R^+)$

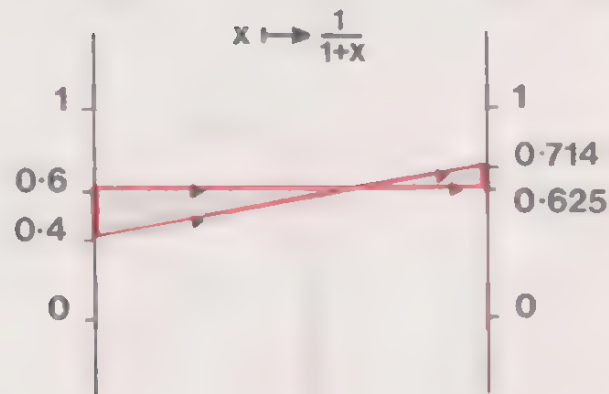(ii) To what does the error interval $[-0.2, 0.2]$ map under the function (a)?

Solution of Example 1

In these cases we can determine the actual bounds directly and need not estimate.

(i) (a) $[0.16, 0.36]$

$$x \longmapsto x^2$$

(b) $[0.625, 0.714]$

$$x \longmapsto \frac{1}{1+x}$$
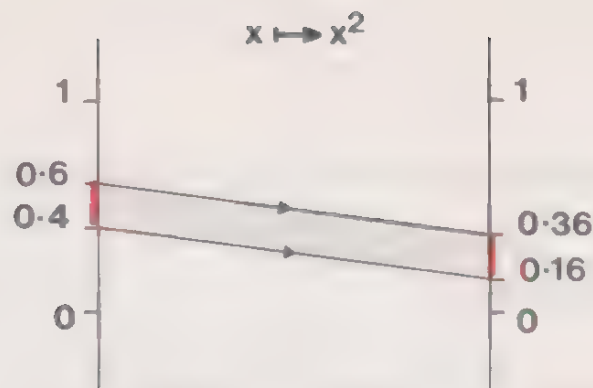
Notice the crossover. The numbers given in the codomain are accurate to three places of decimals.

If we specified the interval by giving its mid-point 0.5 with absolute error bound 0.1 and used the estimating procedure from the preceding section, we would get

absolute error bound in $(1 + x) = 0.1$

relative error bound in $(1 + x) = \dfrac{0.1}{1 + x}$

By the division rule, we have

the estimated relative error bound in

$$\frac{1}{1 + x} = \left| -\frac{0.1}{(1 + x)} \right|$$

so that the estimated absolute error bound in

$$\frac{1}{1 + x} = \left| -\frac{0.1}{(1 + x)^2} \right|$$

This holds for all error intervals of half-width 0.1, but we are interested in the one which is centred on 0.5.

$$\text{Here, estimated absolute error bound} = \frac{0.1}{(1.5)^2} = 0.044$$

$$\text{The image of 0.5 is} \qquad \frac{1}{1.5} = 0.667$$

Thus the estimated error interval in the codomain is

$$[0.667 - 0.044, \ 0.667 + 0.044] = [0.623, 0.711]$$

with end-points differing by only 0.002 or 0.003 from the exact ones calculated above.
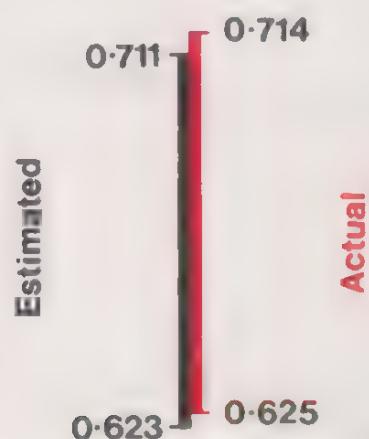
For *particular* error intervals in the domain the estimation method again took longer. Its power lies in its generality as we shall see in the next exercise.

First we must complete the solution of the example.

(ii)
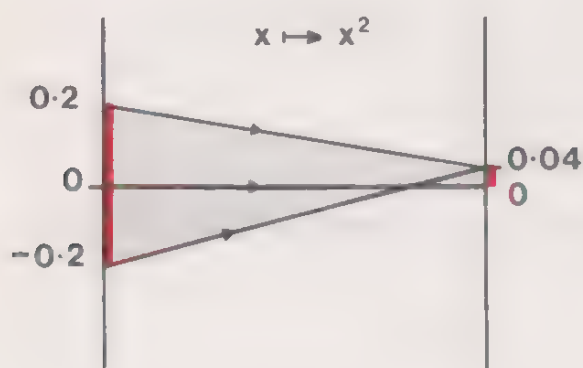
If we adopt the method of estimation given in (i), the interval appears to shrink to nothing: but consider *some other numbers* in the interval. You will see that the lower end-point in the codomain is the image of zero; so beware. If you use this method, always make a quick check of the numbers inside the interval to make sure that their images are behaving themselves.

For the next exercise, we need the following definition.

The scale factor for a function of one variable, propagating an error from the domain to the codomain, is defined as

$$\frac{\text{estimated error in image of } x}{\text{error in } x} \qquad (x \in \text{domain})$$

Note that this definition of the scale factor also gives us an estimate of the ratio

$$\frac{\text{error interval width in codomain}}{\text{error interval width in domain}}$$

since we simply choose two elements, the upper and lower bounds of the interval, in the definition. In other words, if the scale factor is greater than 1 the interval length is magnified, but if it is less than one, the length shrinks. If its sign is negative it implies "crossover", as in the figure in the solution of Example 1(i)(b).

Exercise 1

Using the results of Exercise 7.3.1, calculate the scale factor for the following functions:

(i) $x \longmapsto x^3 - 4x + 3 \quad (x \in R)$

(ii) $x \longmapsto \dfrac{5}{x} - \dfrac{3}{x^2} \qquad (x \in R, x \neq 0)$

(iii) $x \longmapsto \frac{1}{5}(x^3 + 3) \qquad (x \in R)$

at $x = -2, x = \frac{2}{3}, x = 2$.

7.5 Approximating Functions

So far we have considered errors which arise from computations carried out with approximate data: that is, the numbers whose images we calculated are inaccurate. Suppose now that the original data is exact, then in principle, given any x belonging to the domain, it is always possible to find the image of x by applying the rule by which the function is defined. In practice, however, it may not be possible to use this method every time an image under the function is required: the rule may be too complicated to apply conveniently, or in some cases it may not even be known. In such cases it may be that the images (or even approximate values of the images) of some particular numbers $x_1, x_2, x_3, \ldots$, in the domain can be found and in this case these images can be listed together with $x_1, x_2, x_3, \ldots$, in the form of a table. Such a table is a subset of the graph of the function. The most familiar examples of such tables are the tables of logarithms and of trigonometrical functions which we learn to use at school.

x (exact data)	$\log x$ (accurate data, to 4 figures)
1.0	0.000
1.1	0.0414
1.2	0.0792
1.3	0.1139
1.4	0.1461
1.5	0.1761

Table I

We often face the problem of estimating images of values of x that are not tabulated: for example, can we, if necessary, obtain an accurate value of log 1.05, or of log 1.6, from the values given in Table I? This is the type of problem with which we shall deal in this section.

For functions such as the logarithm, sine or cosine, the use of tables is a matter of convenience only: with the aid of a computer we can calculate the logarithm or sine or cosine of a number without using any tables. Many functions, however, are formed from much more complicated rules than the one defining the logarithm: an example would be the function:

$$\left\{ \begin{array}{l} \text{Position on an} \\ \text{aeroplane wing} \end{array} \right\} \longmapsto \left\{ \begin{array}{l} \text{Net upward force on unit area} \\ \text{of the wing at that position} \end{array} \right\}$$

To calculate the images under this function for a given aeroplane wing (under given conditions of flight), even at a few dozen positions on the wing, would be a major computing project, and the more positions were taken the more costly the project would be. It is therefore much more practical to tabulate the images under the function at not too many positions and use the methods of this unit to estimate the images at other positions. Thus the computer revolution has not made the techniques for working with tabulated functions (i.e. functions for which a subset of its graph is specified by a table) obsolete; rather, it has greatly extended the range of functions to which it is possible to apply them.

There are other situations too in which we may come across tabulated functions. The tabulated function may result from a sequence of experimental observations, as for example, in the tables used by engineers for designing steam engines. Here is an extract from such a table*:

Pressure (lbf/in^2)	Boiling Point of Water (°F)
300	417.33
350	431.72
400	444.59
450	456.28

In this case the function is defined by a physical relationship rather than by a mathematical one such as the logarithm, but the domain and co-domain of the function are again subsets of R, and so the method of using the table is just the same as before. An engineer who needed the boiling point of water at a pressure of say, 325 lbf in^2 could estimate it from the table. This problem of estimating the image of a number lying between two of the tabulated numbers in the domain is called the problem of interpolation. The objective of this section is to show you accurate and convenient methods for doing interpolation.

* O. W. Eshbach, *Handbook of Engineering Fundamentals*, (John Wiley 1966).

In order to do an interpolation it is necessary first to find a simple function that has the same image values as the tabulated function at the tabulated elements in the domain.

Since the whole point of the interpolation is to approximate a complicated function by a simpler one, there will be no unique way of choosing the simple function, as we have already seen in Chapter 5.

The following example discusses the above remarks in a little more detail in our present context.

Example 1

A part of a table for the sine function

$$x \longmapsto \sin x \quad (x \in R)$$

is shown below.

x	$\sin x$
-0.4	-0.39
-0.3	-0.30
-0.2	-0.20
-0.1	-0.10
0	0
0.1	0.10
0.2	0.20
0.3	0.30
0.4	0.39

(The values of x are given in radians and the corresponding images are given to two significant figures.)

Now, it would seem reasonably clear that for any element in the restricted domain $[-0.4, 0.4]$, the function $x \longmapsto x$ has approximately the same image as $x \longmapsto \sin x$. Therefore, if we were only interested in this restricted domain for the sine function, we could safely *approximate the complicated function*

$$x \longmapsto \sin x \qquad (x \in [-0.4, 0.4])$$

by the simple function

$$x \longmapsto x \qquad\qquad (x \in [-0.4, 0.4])$$

(This is, of course, just the tangent approximation to sine discussed in Chapter 5.) However, as we said above, this approximation is not unique: for instance, both

$$x \longmapsto \cos\left(\frac{\pi}{2} - x\right) \quad (x \in [-0.4, 0.4])$$

and

$$x \longmapsto x - \frac{x^3}{6} \qquad (x \in [-0.4, 0.4])$$

are also approximations to the sine mapping with this restricted domain (the first happens to be exact, in that $\cos\left(\dfrac{\pi}{2} - x\right) = \sin x$, $(x \in R)$). The non-uniqueness is not something we can get rid of, as you can readily appreciate by considering attaching precise meanings to "complicated", "simple" and "approximate".

As a consequence of the non-uniqueness, there are different methods of interpolation which may well yield different results. There is no absolute criterion for choosing between these methods, except that, when the result is to be applied to, say, an engineering construction, we have a criterion of sorts: does it work? All that can be done is to use whatever information we have about the source of the tabulated numbers and to choose the new function in the simplest way that is likely to yield a good approximate representation of the original function. In order to make a good choice it is important to be able to estimate the error in an interpolation method.

7.6 Linear Interpolation

The simplest method of interpolation is one that is widely used in tables of elementary functions, such as log tables. In the introduction to section 7.5, for example, we mentioned the problem of finding log 1.05 from the following table:

x	$\log x$
1.0	0.0000
1.1	0.0414
1.2	0.0792
1.3	0.1139
1.4	0.1461
1.5	0.1761

Since 1.05 is half way between 1.00 and 1.10, the simplest estimate is to assume that log 1.05 is half way between log 1.00 and log 1.10, that is

$$\log 1.05 \simeq \frac{0.0000 + 0.0414}{2} = 0.0207 \qquad \text{Equation (1)}$$

This agrees to 3 decimal places with the value 0.0212 given in four-figure log tables. (This latter value is itself correct only to 4 decimal places; the exact value of log 1.05 is a non-terminating decimal.) In a similar way, to find log 1.13, since 1.13 is $\frac{3}{10}$ of the way from 1.10 to 1.20, the simplest estimate is log 1.10 plus $\frac{3}{10}$ of the distance to be covered in going from log 1.10 to log 1.20, that is

$$\log 1.13 \simeq \log 1.10 + \tfrac{3}{10}(\log 1.20 - \log 1.10)$$

$$= 0.0414 + \tfrac{3}{10}(0.0792 - 0.0414)$$

$$= 0.0527 \qquad \text{Equation (2)}$$

This agrees to 3 decimal places with the value 0.0531 given in four-figure tables.

The method used in deriving Equations (1) and (2) is sometimes called the method of proportional parts, since the number on the right-hand side of Equation (2) divides the interval [log 1.10, log 1.20] in the same proportion that the number 1.13 divides the interval [1.10, 1.20]. To deal with the general case, let us denote the tabulated function by f and the numbers in the domain for which it is tabulated by $x_1, x_2, \ldots, x_n$, arranged in increasing order. These are known as tabular points. The images of these points are then $f(x_1), f(x_2), \ldots, f(x_n)$; they are called

x		log
1.00		.0000
1.05		←
1.10		.0414
1.13		←
1.20		.0792
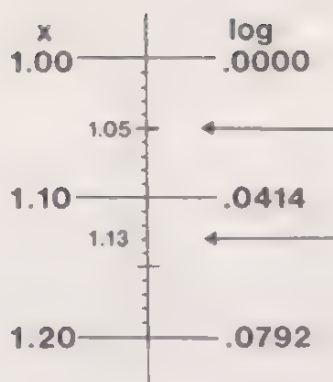
tabular values of the function. For brevity we shall denote the tabular values by $y_1, y_2, \ldots, y_n$:

x	$f(x)$
x_1	$y_1 = f(x_1)$
x_2	$y_2 = f(x_2)$
$\ldots$	$\ldots$
x_n	$y_n = f(x_n)$

To estimate $f(x)$ when x lies between the two tabular points x_1 and x_2 say, we find the proportion in which the number x divides the interval $[x_1, x_2]$ and find the number y that divides the interval $[y_1, y_2]$ in the same proportion. We can represent the procedure graphically:
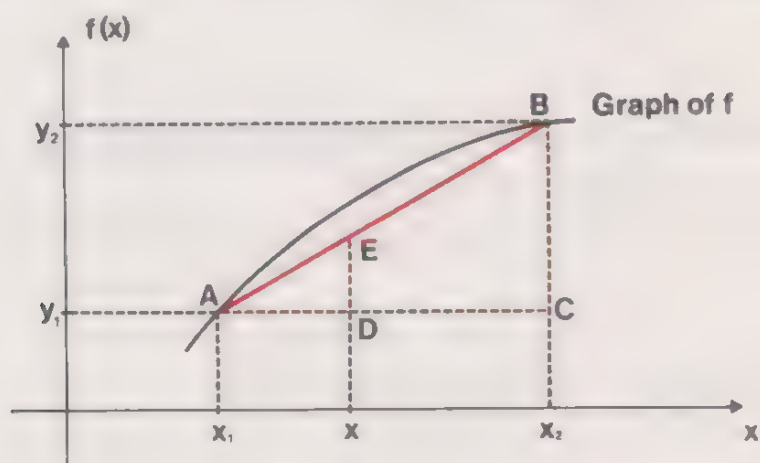
The proportion in which the number x divides the interval $[x_1, x_2]$ is

$$\frac{x - x_1}{x_2 - x_1} = \frac{AD}{AC}$$

Using the fact that triangles ADE and ACB are similar, we have

$$\frac{AD}{AC} = \frac{DE}{CB}$$

Since CB is the length of the interval $[y_1, y_2]$, E is the point whose y-co-ordinate divides the interval $[y_1, y_2]$ in the required proportion. So we see that our method of estimating $f(x)$ between A and B is to approximate its graph by a straight line between the two points. For this reason this method is called linear interpolation.

Some published tables give a little help in the method of linear interpolation by printing the values of $y_{k+1} - y_k$ to the right of the values of y_k and y_{k+1}, and on a line halfway between them, as for example in the following extract from a table of reciprocals:

x	$1/x$	
1.60	0.6250	
		-39
1.61	0.6211	
		-38
1.62	0.6173	
		-38
1.63	0.6135	

The numbers -39, -38, -38, known as first differences, should really be -0.0039, -0.0038, -0.0038, but they are usually printed as shown (or even without the minus sign) to save space.

Exercise 1

Use linear interpolation to obtain approximate values for

(i) log 1.25, using Table I on page 214
(ii) the boiling point of water at a pressure of 365 lbf/in² from the following table

Pressure (lbf/in²)	Boiling Point of Water (°F)
300	417.33
350	431.72
400	444.59
450	456.28

The method of linear interpolation can be applied to any table whatever, but only in suitable cases can we rely on the values so obtained. The criterion is that the graph of the function should be approximately straight between the tabular points.

Exercise 2

Tabulate the values of $\sin x$ for $x = 0$, π, 2π, 3π radians, and calculate a value of $\sin \dfrac{\pi}{10}$ by linear interpolation. What do you conclude?

Exercise 3

(i) Tabulate the values of $\sin x$ for $x = 0, \dfrac{\pi}{4}, \dfrac{\pi}{2}, \dfrac{3\pi}{4}$ and π, and calculate a value for $\sin \dfrac{\pi}{10}$ by linear interpolation.

(ii) Calculate a value for $\sin \dfrac{\pi}{10}$ by linear interpolation from the table at the right.

x	$\sin x$
0	0
$\dfrac{\pi}{16}$	0.1951
$\dfrac{2\pi}{16}$	0.3827
$\dfrac{3\pi}{16}$	0.5555
$\dfrac{4\pi}{16}$	0.7071

(iii) To four places of decimals, the value of $\sin \dfrac{\pi}{10}$ is 0.3090. What do you conclude by comparing this with the results of the preceding exercise and (i) and (ii) of the present exercise?

Exercise 2 illustrates a case where linear interpolation is completely inappropriate : the sine function cannot be satisfactorily represented by a straight line joining tabular points as far apart as 0 and π, as is obvious from the figure.
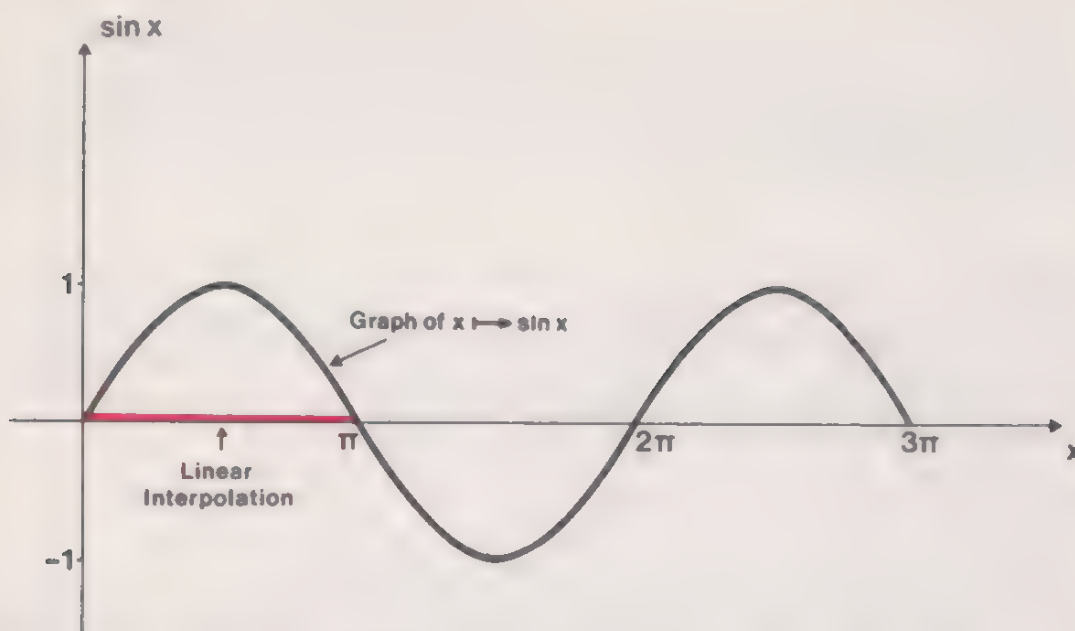
On the other hand, Exercise 3 shows two cases where the tabular points are closer together and where linear interpolation gives fairly good results. In (i) the spacing between tabular points is $\frac{\pi}{4}$, and the interpolation is in error by $0.3090 - 0.2828 = 0.0262$; in (ii) the interval width is only $\frac{\pi}{16}$ and the interpolation is in error by only $0.3090 - 0.3077 = 0.0013$.

Thus, in this example, the accuracy of linear interpolation is generally better, the closer together the tabular points. This is illustrated in the figure.
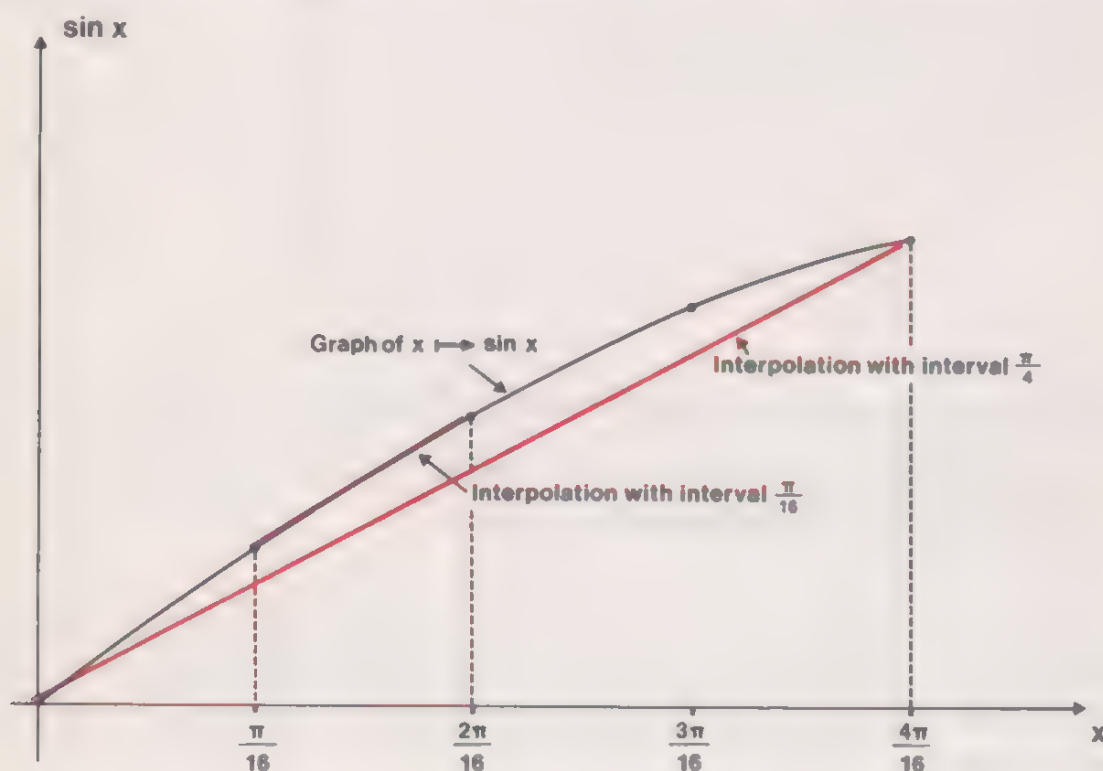
In designing mathematical tables such as log tables the tabular spacing is usually chosen small enough to make the error in linear interpolation no greater than the round-off error in the tabulated figures. This implies that a higher accuracy in the tables demands a smaller tabular spacing and hence more tabular entries. For example, in *Chambers's Four-Figure Tables* the logarithms occupy 4 pages, but in *Chambers's Six-Figure Tables*, (100 times as accurate), they occupy 26 pages. In four-figure tables it is customary to make the interpolation more convenient, at some cost in accuracy, by replacing the proportional parts by mean proportional parts, which are simply proportional parts averaged over a group of tabular intervals (instead of taking the nearest tabular points).

Table of Reciprocals

x	0	1	2	3	4	5	6	7	8	9
16	6250	6211	6173	6135	6098	6061	6024	5988	5952	5917
17	5882	...								

Mean Proportional Parts

1	2	3	4	5	6	7	8	9
4	7	11	15	18	22	26	30	33

For example, in the above table of reciprocals, the mean proportional part under the number 2 in the table is obtained as

$$\frac{(5882 - 6250)}{10} \times \frac{2}{10} = -\frac{736}{100} \simeq -7$$

(The minus sign is not shown in the table for brevity; you are expected to remember to subtract, since $\frac{1}{x}$ decreases as x increases.)

So whether you are reading the reciprocal of 16.12 or of 16.72 you use the same mean proportional part, i.e. -7. Whereas, if you were to use the nearest tabulated points, you would use linear interpolation on the tabulated values at 16.1 and 16.2 for 16.12, obtaining the true proportional part, $0.2(6173 - 6211) = -8$, rather than -7, and on the tabulated values at 16.7 and 16.8 for 16.72, obtaining the true proportional part $0.2(5952 - 5988) = -7$.

Exercise 4

Using the table of reciprocals, read off the reciprocals of 1.66 and 1.67 (in such tables you are often expected to find the position of the decimal point for yourself). By linear interpolation, obtain a value for the reciprocal of 1.667. Also obtain a value for the reciprocal using the mean proportional parts shown, and explain any discrepancy.

Summary

In this section we have shown that the images of a tabulated function, corresponding to values between known tabular values, can be conveniently calculated using linear interpolation, provided that within the interval being used for interpolation the graph of the function is reasonably close to a straight line.

7.7   Polynomial Interpolation

Earlier in this chapter we used an extract from four-figure logarithm tables. Such tables exist for various numbers of figures, five, six and seven being quite common and, as we have remarked, in tables in common use, the tabular spacing is small enough to make linear interpolation accurate and convenient. For a table that will not see so much use, however, the designer may well choose to economize by computing the value of the function for relatively few tabular points; he is particularly likely to do this if each tabular value demands a lot of computing effort and if the resulting table will be used only once because it arises from a unique situation, as in the case of the function mentioned in section 7.5, which describes the lifting effect of the air on particular parts of an aeroplane wing under particular flying conditions. In such cases the tabular spacing is likely to be too large to justify linear interpolation and more sophisticated methods are necessary.

The purpose of this section is to show you one such method, in which the tabulated function is approximated between the tabular points not by a linear function but by a polynomial function (cf. Chapter 5). You may ask: "What is special about polynomial functions; why not use some other function such as the sine or cosine?" One reason is simply that polynomial interpolation is very convenient; and convenience is a very important factor in making a choice of numerical methods.

A further reason for using polynomials is contained in a theorem proved by a German mathematician, Weierstrass, who lived in the last century. This theorem (whose proof we shall not consider here) states that any continuous function can be approximated over any desired interval to any desired accuracy by means of some polynomial. Unfortunately Weierstrass' theorem has the annoying feature of so many so-called *existence theorems* in mathematics: it tells us that the polynomial exists, but not how to find it. In the rest of this section we shall not use Weierstrass' theorem except to give confidence in the ultimate validity of polynomial approximation methods, and concentrate instead on practical methods for finding approximating polynomials and for interpolating with their help.

The very simplest type of polynomial interpolation is the same thing as linear interpolation, which we have already considered. The linear functions used there to approximate the tabulated functions are none other than polynomial functions of degree 1. Before discussing non-linear interpolation, therefore, we recapitulate the principles of linear interpolation using a slightly different viewpoint from the one used in the section devoted to this topic.

In linear interpolation we approximate the function between two successive tabular points, say x_k and x_{k+1}, by a linear function that fits (i.e. has the same tabular values as) the tabulated function at those two tabular points. (See figure.)
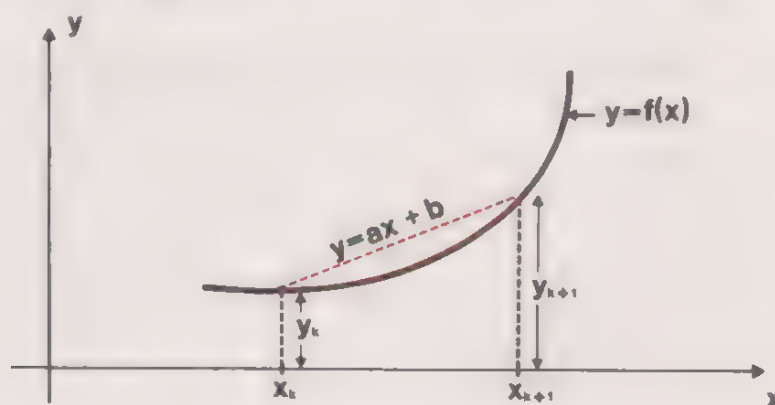
This linear function may be denoted by l, where

$$l : x \longmapsto ax + b \qquad (x \in [x_k, x_{k+1}])$$

and a and b are two real numbers.

Exercise 1

Why did we use $[x_k, x_{k+1}]$, rather than R, for the domain of l?

We have stipulated that l must fit the tabulated function at x_k and x_{k+1}; that is,

$$
\left.\begin{array}{c}
l(x_k) = f(x_k) \\
l(x_{k+1}) = f(x_{k+1})
\end{array}\right\}
$$

These two conditions are just enough to determine the two constants a and b in the definition of l; for, they are equivalent to

$$
\left.\begin{array}{c}
x_k a + b = f(x_k) \\
x_{k+1} a + b = f(x_{k+1})
\end{array}\right\}
$$

and since the tabular points x_k, x_{k+1} and the values $f(x_k)$, $f(x_{k+1})$ are known, we can treat these equations as a pair of simultaneous equations for the two unknowns a and b.

The solution of the simultaneous equations is

$$
\left.\begin{array}{l}
a = \dfrac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} \\[3mm]
b = \dfrac{x_{k+1} f(x_k) - x_k f(x_{k+1})}{x_{k+1} - x_k}
\end{array}\right\}
$$

which when substituted into the formula for $l(x)$ gives

$$
l(x) = \frac{x_{k+1} - x}{x_{k+1} - x_k} f(x_k) + \frac{x - x_k}{x_{k+1} - x_k} f(x_{k+1}) \qquad \text{Equation (1)}
$$

We shall carry this solution a little further and eliminate x in favour of the variable

$$
\theta = \frac{x - x_k}{x_{k+1} - x_k}
$$

which represents the proportion in which the number x divides the interval $[x_k, x_{k+1}]$ (see figure).

Noticing that

$$
\frac{x_{k+1} - x}{x_{k+1} - x_k} = 1 - \frac{x - x_k}{x_{k+1} - x_k} = 1 - \theta
$$

Equation (1) now becomes

$$l(x) = f(x_k) + \{f(x_{k+1}) - f(x_k)\}\theta, \quad (x \in [x_k, x_{k+1}]) \qquad \text{Equation (2)}$$

A neater way to write Equation (2) is to use the difference operator*: it then simplifies to

$$l(x) = f(x_k) + \theta \cdot \Delta_h f(x_k) \quad (x \in [x_k, x_{k+1}]) \qquad \text{Equation (3)}$$

Exercise 2

Use Equation (3) to estimate $\tan(1.444)$ from the table at the right, and compare with the true value (to three decimal places), $\tan(1.444) = 7.844$.

x	$\tan x$
1.43	7.055
1.44	7.602
1.45	8.238
1.46	8.989

(You could, of course, answer this exercise by the methods of section 7.6, but in order to get a feeling for our present methods, we suggest you do it using Equation (3).)

When linear interpolation is not very accurate, as in the preceding exercise, it indicates that the tabulated function is not very accurately represented by the linear function $l(x)$ over the interval $[x_k, x_{k+1}]$. In such cases we can try to allow for the non-linearity by using a simple non-linear approximating function instead of a linear one. If the approximating function is to be a polynomial, then the simplest (i.e. of lowest degree) non-linear possibility is the quadratic function

$$q: x \longmapsto ax^2 + bx + c$$

To complete the specification of this function we need values for the three constants a, b, c, and also a specification of the domain. Following the method used in the preceding section we could try to determine a, b, and c by requiring that q must fit the tabulated function at two successive

---

* The difference operator was defined in Volume 1, Chapter 8 by

$$\Delta_h: f \longmapsto [x \longmapsto f(x+h) - f(x)]$$

or in terms of image by

$$\Delta_h f(x) = f(x+h) - f(x)$$

tabular points, say x_k and x_{k+1}. Can you see the difficulty that would arise? We would get two simultaneous equations

$$\left.\begin{array}{c} x_k^2 a + x_k b + c = f(x_k) \\[6pt] x_{k+1}^2 a + x_{k+1} b + c = f(x_{k+1}) \end{array}\right\}$$

but they would not be enough to determine a, b, and c (see figure).
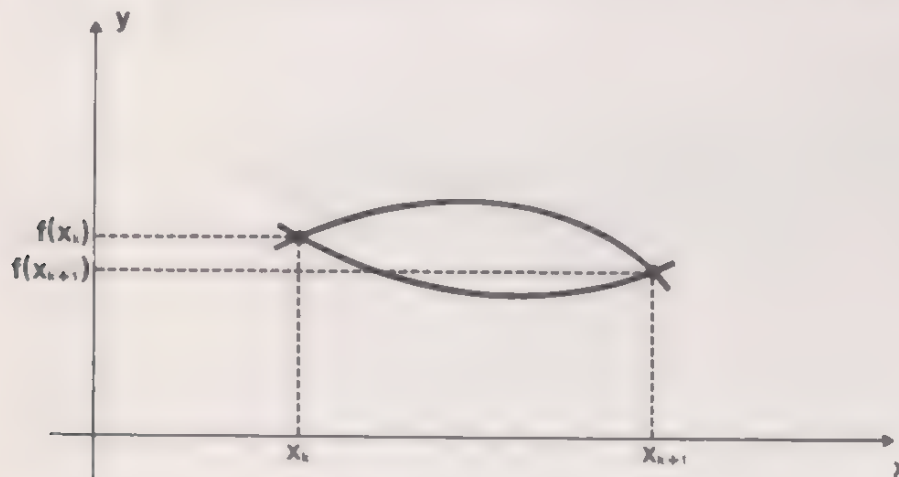
To find three unknown quantities we need *three* simultaneous equations. There are various ways of obtaining a third equation; the simplest, and the only one we shall consider here, is to require that q shall fit the tabulated function at *three* consecutive tabular points instead of just two (see figure).

Graph of one of the many quadratics fitting the tabulated function at only 2 tabular points.

Graph of the unique quadratic fitting the tabulated function at the 3 consecutive tabular points.
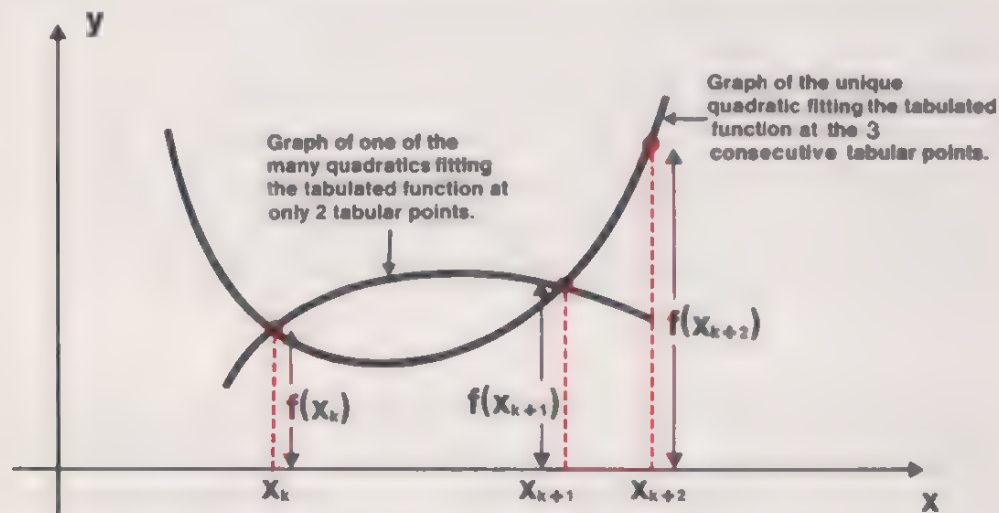
Calling these three points x_k, x_{k+1}, x_{k+2}, we then have the three equations

$$\left.\begin{array}{l} q(x_k) = x_k^2 a + x_k b + c = f(x_k) \\[6pt] q(x_{k+1}) = x_{k+1}^2 a + x_{k+1} b + c = f(x_{k+1}) \\[6pt] q(x_{k+2}) = x_{k+2}^2 a + x_{k+2} b + c = f(x_{k+2}) \end{array}\right\}$$ **Equations (4)**

and since the number of equations is now equal to the number of un-knowns, the equations now contain enough information to determine a, b, and c.

It remains to specify the domain of $q(x)$. Because of Equations (4) the domain must include at least the numbers x_k, x_{k+1}, and x_{k+2}, and, if the function is to be useful for interpolation, it must also include intermediate values; thus the least domain for $q(x)$ is $[x_k, x_{k+2}]$. The graph of this function

$$q : x \longmapsto ax^2 + bx + c \quad (x \in [x_k, x_{k+2}])$$

is the one shown in the last figure. Of course, if the function were required (and were suitable) for extrapolation, the domain could easily be extended.

To find the three constants a, b, c, the obvious method of proceeding is to solve the three simultaneous equations (4).

There are standard techniques for solving such a system of equations; however, we are not concerned with the techniques but only with the final result, which can be put in the form

$$
\begin{aligned}
q(x) &= ax^2 + bx + c \\
&= \frac{(x - x_{k+1})(x - x_{k+2})}{(x_k - x_{k+1})(x_k - x_{k+2})} f(x_k) \\
&\quad + \frac{(x - x_k)(x - x_{k+2})}{(x_{k+1} - x_k)(x_{k+1} - x_{k+2})} f(x_{k+1}) \\
&\quad + \frac{(x - x_k)(x - x_{k+1})}{(x_{k+2} - x_k)(x_{k+2} - x_{k+1})} f(x_{k+2}) \qquad \text{Equation (5)}
\end{aligned}
$$

The expression on the right is called Lagrange's interpolation polynomial; it is the analogue of Equation (1) for the linear case. (There is no point in your learning this formula by heart.)

Just as in our treatment of the linear case we can simplify Equation (5) but we will not bother.

The formula for $q(x)$ is a special case of a general formula called the Gregory-Newton interpolation formula.

Exercise 3

Use the above quadratic interpolation formula to calculate tan (1.444) from the table at the right, and compare with the result obtained by linear interpolation in Exercise 2, and with the correct result 7.844.

x	$\tan x$
1.43	7.055
1.44	7.602
1.45	8.238
1.46	8.989

7.8   Answers to Exercises

Section 7.1

Exercise 1

$$\frac{8}{0.4} \times \frac{8}{0.4} \times \frac{10}{0.4} = 10\,000$$

It is quite surprising that the 20% change in the supposed diameter of the pea nearly doubles the estimate of the number.

Section 7.2

Exercise 1

The approximate value x is 2.5 kg.

(i) Maximum magnitude of e_x is 0.05 kg = absolute error bound.
(ii) Maximum magnitude of the relative error is

$$\frac{0.05\,\text{kg}}{2.5\,\text{kg}} = 0.02 = \text{relative error bound}.$$

Exercise 2

0.005 in all cases, which is 0.05%, 5% and 50% percentage relative error respectively.

Exercise 3

	ε_x	ρ_x
(i)	1 sec	5×10^{-11}
(ii)	30Ω	0.10
(iii)	$2\mu F$	0.20

Exercise 4

NO. The answer quoted in the question is $\sqrt{5}$ correct to five places of decimals, but this is meaningless in the context of the question, since this implies five-figure accuracy in the hypotenuse, whereas the original data had an accuracy only to the nearest centimetre. A reasonable answer would be 2.24 metres.

Section 7.3

Exercise 1

(i) Error in the image $= e_{x^3} - 4e_x + 0$. Therefore, estimated error in the image $= (3x^2 - 4)e_x$. (Notice that we have used the fact that $e_{-4x} = -4e_x$.)

(ii) We have already found the error estimates

$$e_{1/x} \simeq -\frac{e_x}{x^2}, \qquad e_{1/x^2} \simeq -\frac{2e_x}{x^3}$$

so that the error estimate in

$$\left(\frac{5}{x} - \frac{3}{x^2}\right) = -\frac{5e_x}{x^2} + \frac{6e_x}{x^3} = \left(\frac{6 - 5x}{x^3}\right)e_x$$

Section 7.4

Exercise 1

(i) We found in Exercise 7.3.1(i) that the estimated error in the image of $(x + e_x)$ was $(3x^2 - 4)e_x$.

Therefore the scale factor is

$$\frac{(3x^2 - 4)e_x}{e_x} = 3x^2 - 4$$

x	scale factor
$+2$	8
$\frac{2}{3}$	$-\frac{8}{3}$
-2	8

$$x \longmapsto x^3 - 4x + 3$$
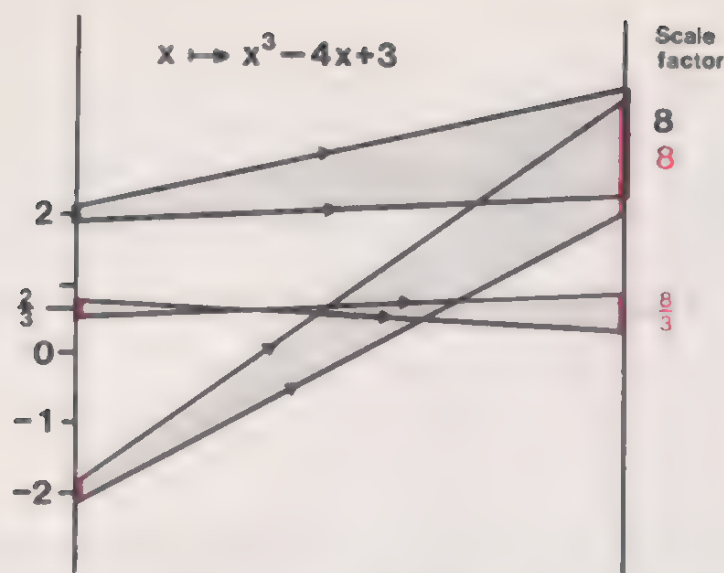
The diagram shows how the error intervals in the domain are magnified in the codomain.

(ii) Using the result from Exercise 7.3.1(ii), the scale factor is

$$\frac{6 - 5x}{x^3} \quad (x \neq 0)$$

$$x \longmapsto \frac{5}{x} - \frac{3}{x^2}$$
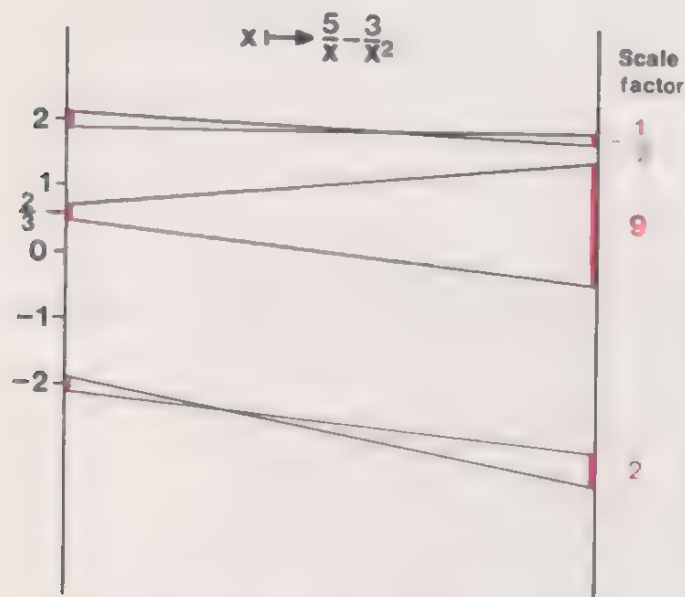
x	scale factor
$+2$	$-\frac{1}{2}$
$\frac{2}{3}$	9
-2	-2

Here, the two error intervals centred on $\frac{2}{3}$ and -2 in the domain grow, but the one centred on 2 shrinks.

(iii) In this case the scale factor is $\dfrac{3x^2}{5}$

$$x \longmapsto \tfrac{1}{5}(x^3+3)$$
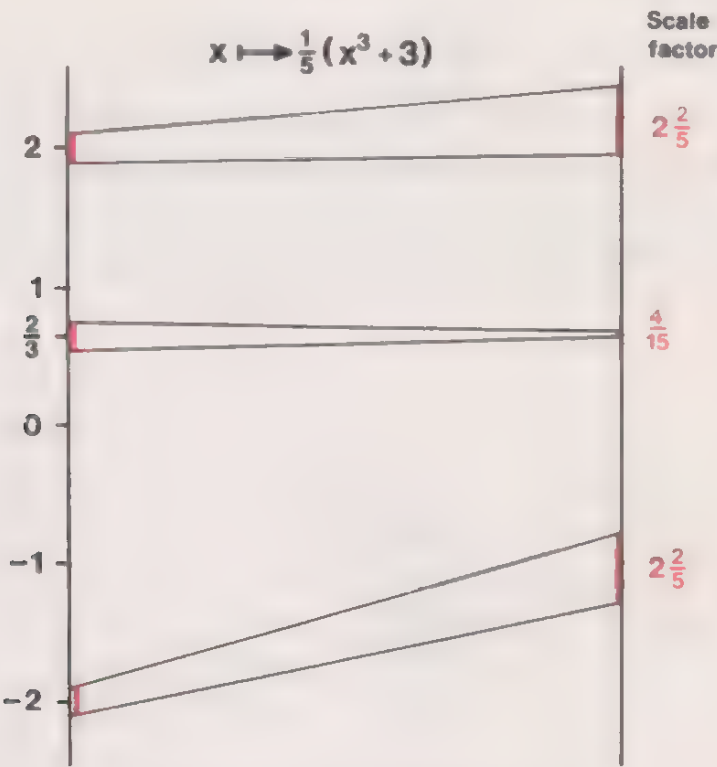
x	scale factor
$+2$	$\frac{12}{5}$
$\frac{2}{3}$	$\frac{4}{15}$
-2	$\frac{12}{5}$

Here, the two error intervals centred on 2 and -2 in the domain grow, but the one centred on $\frac{2}{3}$ shrinks. No crossover occurs, since all three scale factors are positive.

Section 7.6

Exercise 1

(i) 0.0966
(ii) 435.58 °F.

Exercise 2

x	0	π	2π	3π
$\sin x$	0	0	0	0

Linear interpolation gives $\sin \dfrac{\pi}{10} = 0$. Indeed, linear interpolation gives $\sin x = 0$ for all values of x between 0 and 3π. The conclusion is that linear interpolation can give wildly inaccurate results if the interval of tabulation is too large. One may conclude also that extreme care must be taken in cases where little is known about the function being tabulated.

Exercise 3

(i)

x	$\sin x$
0	0.0000
$\dfrac{\pi}{4}$	0.7071
$\dfrac{\pi}{2}$	1.0000
$\dfrac{3\pi}{4}$	0.7071
π	0.0000

The interpolated value for $x = \dfrac{\pi}{10}$ is $0.0000 + (0.7071 - 0.0000)\theta$, with

$$\theta = \frac{\dfrac{\pi}{10} - 0}{\dfrac{\pi}{4} - 0} = \frac{4}{10},$$

$$= 0.0000 + 0.7071 \times \frac{4}{10} = 0.2828$$

(ii) The interpolated value for $x = \dfrac{\pi}{10}$ is

$$\sin\left(\frac{\pi}{16}\right) + \left[\sin\left(2\frac{\pi}{16}\right) - \sin\left(\frac{\pi}{16}\right)\right]\theta$$

with

$$\theta = \frac{\dfrac{\pi}{10} - \dfrac{\pi}{16}}{\dfrac{2\pi}{16} - \dfrac{\pi}{16}} = \frac{6}{10}$$

$$= 0.1951 + \left[0.1876 \times \frac{6}{10}\right] = 0.3077$$

(iii) The interpolated values get closer to the given four-figure value as we reduce the interval between the tabular points.

Exercise 4

$$1/1.66 = 0.6024, \quad 1/1.67 = 0.5988$$

Linear interpolation

$$1/1.667 \simeq 0.6024 + (0.5988 - 0.6024)\theta$$

with

$$\theta = \frac{1.667 - 1.66}{1.67 - 1.66} = \frac{7}{10}$$

so that $1/1.667 \simeq 0.6024 - 0.0036 \times \dfrac{7}{10} = 0.5999$

Mean proportional parts

$$1/1.667 \simeq 0.6024 - 0.0026 = 0.5998$$

(Mean proportional part for $\theta = \frac{7}{10}$ is shown in the table as 26.)

The results obtained using mean proportional parts are slightly inaccurate because the proportional parts for $\frac{7}{10}$ vary between

$$(6250 - 6211) \times \tfrac{7}{10} = 27$$

and

$$(5952 - 5917) \times \tfrac{7}{10} = 24$$

across the table, so that using the mean value 26 can introduce an error of up to 2 digits in the last decimal place.

Section 7.7

Exercise 1

The graph of *l* is to be a segment of a straight line : not a complete straight line. The algebraic equivalent of this geometric condition is the restriction on the domain of *l*. Evidently, we wish to approximate to the complete curve like this :

not like this :

Exercise 2

The interval $[x_k, x_{k+1}]$ containing 1.444 is [1.44, 1.45]; hence $x_k = 1.44$ and

$$\theta = \frac{0.004}{0.01} = \frac{4}{10}$$

$$\Delta_h f(x_k) = f(x_{k+1}) - f(x_k) \quad \text{becomes}$$

$$\Delta_{0\cdot01} f(1.44) = f(1.45) - f(1.44) = 0.636.$$

Equation (3) thus gives

$$\tan(1.444) \simeq 7.602 + \tfrac{4}{10} \times 0.636 = 7.856$$

which is the estimate for $\tan(1.444)$ given by linear interpolation. The error of 0.012 shows that the approximation of $\tan x$ by $l(x)$ does not give good accuracy in this range.

Exercise 3

The number 1.444 lies in the interval $[x_k, x_{k+2}]$ with $x_k = 1.43$ or 1.44; either value for x_k may be used. With $x_k = 1.43$, we get

$$q(1.444) = -0.12f(1.43) + 0.84f(1.44) + 0.28f(1.45) = 7.846$$

With $x_k = 1.44$, we get

$$q(1.444) = 0.48f(1.44) + 0.64f(1.45) - 0.12f(1.46) = 7.843$$

Choosing $x_x = 1.43$, our solution agrees within 0.002 with the correct value. When x_k is chosen to be 1.44, the agreement is exact to 3 decimal places. Using linear interpolation, we found $\tan(1.444) = 7.856$, which is not very accurate.

Index

A reference in **bold type** indicates that a definition of the term appears on that page. Concepts which are introduced in Volume 1 and are not redefined in this volume are not indexed.

Index

Index

Index

AN INTRODUCTION TO CALCULUS AND ALGEBRA

Volume 2 Calculus Applied

This second volume develops calculus from the background which was prepared in Volume 1. It begins with the *techniques of differentiation* and *integration*, and goes on to introduce *Taylor's theorem* and *differential equations*. The last of its seven chapters is an introduction to some simple concepts in numerical calculation.

This is the second of three volumes presenting some of the essential concepts of mathematics, a few important proofs (usually in outline), together with exercises designed to reinforce the understanding of the concepts and develop the beginnings of technical skill.

A particular feature of these volumes is the use of two-colour printing to emphasize important concepts and definitions and to heighten the impact of diagrams. Exercises are provided throughout, and detailed answers (with additional comment) are given at the end of each chapter.

The selection of material has been made with the needs of students of other subjects particularly in mind. The material presented is, therefore, suited to a wide class of reader and will provide both a modern intro- duction and a handy reference to two important areas of mathematics: Calculus and Algebra.

Volume 1 Background to Calculus
Volume 2 Calculus Applied
Volume 3 Algebra

THE OPEN UNIVERSITY PRESS